Verbeterde onlinediensten door gepersonaliseerde aanbevelingen en optimale parameters voor de gebruikerservaring

Improved Online Services by Personalized Recommendations and Optimal Quality of Experience Parameters

Toon De Pessemier

Promotor: prof. dr. ir. L. Martens
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen: Computerwetenschappen

Vakgroep Informatietechnologie
Voorzitter: prof. dr. ir. D. De Zutter
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2012 - 2013

UNIVERSITEIT GENT

ISBN 978-90-8578-607-8
NUR 984, 986
Wettelijk depot: D/2013/10.500/40

Universiteit Gent
Faculteit Ingenieurswetenschappen en Architectuur
Vakgroep Informatietechnologie

Promotor: Prof. Dr. Ir. Luc Martens

Universiteit Gent
Faculteit Ingenieurswetenschappen en Architectuur

Vakgroep Informatietechnologie
Gaston Crommenlaan 8 bus 201, B-9050 Gent, België

Tel.: +32-9-331.49.08
Fax.: +32-9-331.48.99

Dit werk kwam tot stand in het kader van een
specialisatiebeurs van het FWO-Vlaanderen
(Het Fonds Wetenschappelijk Onderzoek-Vlaanderen)

Proefschrift tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen:
Computerwetenschappen
Academiejaar 2012-2013

Dankwoord

Dit dankwoord kan door u, beste lezer, gezien worden als een toegankelijke inleiding op dit boek boordevol technische informatie over mijn doctoraatsonderzoek. Voor mij betekent dit dankwoord het einde van mijn doctoraat, een hoofdstuk in mijn leven dat afgelopen is, hoewel wetenschappelijk onderzoek eigenlijk nooit klaar is. Op de voorpagina van dit boek staat enkel mijn naam bij de titel vermeld. Een doctoraat komt echter niet tot stand zonder de hulp en de steun van vele personen, die allen rechtstreeks of onrechtstreeks hun bijdrage aan het doctoraat leveren.

Uiteraard wil ik mijn grote dank betuigen aan mijn promotor Luc Martens, die mij de kans gaf om dit onderzoek aan te vatten bij de WiCa onderzoeksgroep. Zijn advies, begeleiding, en jarenlange ervaring hielpen mij bij het uitstippelen van dit onderzoek. Maar hij gaf ook de nodige ruimte en vrijheid om mijn onderzoek volledig te kunnen ontplooien. Ik kon steeds bij hem terecht voor het bediscussiëren van onderzoeksresultaten, en kritische maar opbouwende feedback. Tevens het vernoemen waard zijn het grote aantal pagina's van rapporten, artikels, en dit doctoraatsboek die hij heeft moeten doorworstelen; steeds was hij bereid om deze na te lezen en te corrigeren. Speciale dank gaat ook naar Wout Joseph van onze onderzoeksgroep, aan wie ik steeds raad en feedback over mijn onderzoek kon vragen. Daarnaast dank ik ook de leden van de examencommissies voor hun suggesties ter verbetering van dit doctoraatsboek. Universiteit Gent, iMinds en het FWO-Vlaanderen, bedank ik voor de onderzoekskansen binnen projecten en de financiële ondersteuning van mijn onderzoek.

Mijn collega's van de WiCa onderzoeksgroep wil ik bedanken voor de aangename werksfeer, Aliou, Amine, Arno, David, Emmeric, Frederic, Günter, Kris, Leen, Margot, Marina, Mostafa, Ning, Sam, en Simon. Sommigen van hen verdienen echter nog een speciale vermelding voor hun geleverde inspanningen. Bedankt Kris, voor alle kleine en grote programmeerwerkjes. Simon wens ik te bedanken voor de samenwerking bij het schrijven van artikels, het OMUS project, en de gebruikersevaluatie op de website "Uit in Vlaanderen". Voor advies omtrent een correcte statistische verwerking van de data kon ik steeds terecht bij Emmeric. Bedankt hiervoor! Isabelle, bedankt voor de administratieve ondersteuning bij onkosten, bestellingen en reisaanvragen.

Ook de collega's die mij vergezelden in de beginjaren van mijn doctoraat, maar ondertussen andere oorden hebben opgezocht, wil ik bedanken, Adrian, Damiano, Dyvia, Francis, István, Jeffrey, Lies, Michiel en Tom. Bedankt Adrian en István

voor de samenwerking bij het Gr@sp project. Voor advies rond technologiekeuzes en onderzoekspistes, nog een extra dankjewel voor Tom.

Ook vele mensen van buiten onze onderzoeksgroep hebben een noemenswaardige bijdrage geleverd aan dit doctoraat. De onderzoeksgroep MICT, onder leiding van Lieven De Marez, wil ik bedanken voor de interdisciplinaire samenwerking die noodzakelijk was voor een groot gedeelte van mijn onderzoek. In het bijzonder waardeer ik de enorme hulp die ik gekregen heb van Katrien bij het plannen van experimenten, het uitvoeren van gebruikerstesten, en het schrijven van artikels. Haar expertise en ervaring in het domein van gebruikersonderzoek vormden een belangrijke meerwaarde voor dit onderzoek. Zonder haar was dit onderzoek niet mogelijk, een dikke merci! Bij het onderzoek rond PersonalTV, kon ik rekenen op de hulp van Cédric en Peter voor de statistische verwerking van de resultaten. Sam en Erik van de Multimedia Lab onderzoeksgroep wil ik bedanken voor de vruchtbare samenwerking bij het CUPID en Stream Store project en het schrijven van artikels.

Net zoals voor alle belangrijke (maar ook banale) zaken in het dagdagelijkse leven, kon ik ook voor dit doctoraat altijd rekenen op de steun en herhaaldelijke aanmoedigingen van mijn ouders en zus. Ook mijn vrouw Stephanie heeft steeds mijn vreugdekreten of klachten moeten aanhoren als ik 's avonds thuis kwam na een geslaagde of soms minder geslaagde dag van experimenteren. Bedankt lieve schat voor alle mooie momenten die we reeds beleefd hebben en die nog moeten komen met onze kleine spruit!

Ik kon ook altijd rekenen op de hulp van mijn schoonouders, wie ik dagelijks bestookte met vragen over een ander groot hoofdstuk in mijn leven, namelijk de bouw van ons huis. Hoewel ze dit nog niet kunnen lezen, wil ik Audrey en Jonathan bedanken voor de leuke momenten 's morgens aan de ontbijttafel waardoor mijn werkdag telkens goed van start kon gaan. Ook mijn schoonbroer Sven en schoonzussen Sophie en Leen mogen hier niet ontbreken. Met jullie is het altijd gezellig op familiefeesten, wat mij hielp om mijn gedachten even te verzetten.

Ook al was het niet altijd eenvoudig om in enkele zinnen uit te leggen waarover dit doctoraat precies gaat, of wat de relevatie is voor de man in de straat, de interesse van velen motiveerde me telkens opnieuw voor dit onderzoek. Mijn vrienden hebben elk op hun manier mee dit werk helpen verwezenlijken en dit meestal door het creëren van ontspanningsmogelijkheden.

Bedankt aan de mannen van onze wielerploeg en gelijknamige quizploeg, WTC Chasse Patat: Kevin, Samuël, Stefaan, Steven, Tim en Thomas. Ik kijk al uit naar ons volgend mannenweekend!

Op zondag zorgde voetbal de afgelopen jaren voor de nodige ontspanning. Elke (thuis)match waren ze op post, de "lads" van de Spionkop, supportersclub van "den Iendracht" met de trouwste supporters van't land: Björn, Dieter, Geert, Gio, Jean, Josken, Kristin, Mattis, Pieter, Steve, en Yves. Bedankt !

Zeker ook niet te vergeten zijn de kameraden waarmee geregeld een pintje gedronken wordt en jaarlijks carnaval mee wordt gevierd: Annelien, Bert, Evelien, Joris, Joyce, Kevin, Koen, Maarten D.B., Maarten S., Pieter, Stijn B., Stijn T., Tom, en ik zal er nog wel enkele vergeten zijn. "Weir doeng voesj!"

De iets rustigere zaterdagavonden werden opgevuld met een gezellige babbel of een bordspel samen met Andres, David, Ilse, Inge, Kathleen, Kevin, Kristien, Liesje, Rube, Seppe, en Stijn. Bedankt allemaal !

Gent, juni 2013
Toon De Pessemier

Table of Contents

List of Figures

List of Tables

List of Symbols and Acronyms

A

AAC LC 3 Advanced Audio Coding, Low Complexity profile 3

ACR Absolute Category Rating

ADP1 Android Developer Phone 1

AJ mean Audio Jitter

AL Audio packet-Loss rate

ANCOVA ANalysis of COVAriance

ANOVA ANalysis Of VAriance

API Application Programming Interface

AVC Advanced Video Coding

Avg Average

AvgWM Average Without Misery

C

CB Content-Based

CF Collaborative Filtering

D

DASH Dynamic Adaptive Streaming over HTTP

DCG Discounted Cumulative Gain

DSIS Double Stimulus Impairment Scale

E

EDGE Enhanced Data rates for GSM Evolution

F

FP False Positive

FPR False Positive Rate

FN False Negative

G

GP GPRS Percentage

GPRS General Packet Radio Service

GPS Global Positioning System

GSM Global System for Mobile communication

H

H_0 The null Hypothesis

HSPA High Speed Packet Access

HTTP HyperText Transfer Protocol

I

IR Information Retrieval

IBCF Item-Based Collaborative Filtering

IEEE Institute of Electrical and Electronics Engineers

IP Internet Protocol

IP-TV Internet Protocol TeleVision

ISO International Organization for Standardization

ITU-T International Telecommunication Union - Telecommunication Standardization Sector

L

LM Least Misery

LT Loading Time

M

MAE Mean Absolute Error

MANCOVA Multivariate ANalysis of COVAriance

MANOVA Multivariate ANalysis Of VAriance

MDI Media Delivery Index

MICT Media and ICT

MMLab MultiMedia Lab

MOS Mean Opinion Score

MP Most Pleasure

MPQM Moving Pictures Quality Metric

MTBF Mean Time Between Failures

MTTR Mean Time To Repair or Mean Time To Recovery

N

NN Nearest Neighbourhood

nDCG normalized Discounted Cumulative Gain

NDPM Normalized Distance-based Performance Measure

O

OSI Open Systems Interconnection

P

PL Packet Loss

POI Point-Of-Interest

PSNR Peak Signal-to-Noise Ratio

Q

QoE Quality of Experience

QoS Quality of Service

R

R^2 coefficient of determination

RMSE Root Mean Square Error

ROC Receiver Operating Characteristic

RSSI Received Signal Strength Indication

RTP Real-time Transport Protocol

RTT Round-Trip delay Time

S

SLA Service-Level Agreement

SR Star-Rating

SSIM Structural SIMilarity

Std Standard deviation

SVD Singular Value Decomposition

T

TCP Transmission Control Protocol

TF-IDF Term Frequency-Inverse Document Frequency

TP True Positive

TPR True Positive Rate

TN True Negative

U

UBCF User-Based Collaborative Filtering

UDP User Datagram Protocol

UMTS Universal Mobile Telecommunications System

V

VJ mean Video Jitter

VL Video packet-Loss rate

W

WCDMA Wideband Code Division Multiple Access

WiCa Wireless and Cable

Nederlandstalige samenvatting

Door de opmars van het internet in het laatste decenium, schieten nieuwe onlinediensten als paddenstoelen uit de grond. Denk hierbij maar aan onlinewinkels, informatieve webpagina's, videodiensten, enz. Als gevolg hiervan krijgen gebruikers te kampen met het probleem van overaanbod: hoewel een overvloed aan informatie beschikbaar is, is het dikwijls een moeilijke opdracht om de meest nuttige en relevante informatie terug te vinden. Dit probleem kan verholpen worden met aanbevelingssystemen, die gebruikers helpen bij het ontdekken en selecteren van de meest interessante informatie of items. Aanbevelingssystemen geven gebruikers persoonlijke suggesties (ook aanbevelingen genoemd) op basis van hun voorkeuren en optimaliseren op die manier het nut en de gebruiksvriendelijkheid van onlinediensten. Het eerste deel van dit doctoraatsproefschrift focust op aanbevelingssystemen en hun evaluatie op basis van verschillende kwalitatieve aspecten.

Maar het inhoudelijke (d.w.z. de beschikbare items of informatie) en de functionaliteit zijn niet langer de enige aspecten waarop onlinediensten zich differentiëren. De ervaring die een gebruiker heeft bij zo'n onlinedienst wordt een belangrijke factor in het ontwerp, de ontwikkeling, en het optimalisatieproces van de dienst. Het kwantificeren van de gebruikerservaring of *Quality of Experience* (QoE) blijft echter een uitdaging, zeker voor mobiele media, die gebruik maken van een grote verscheidenheid aan technologieën en netwerken, en de afgelopen jaren een exponentiële groei gekend hebben van mobiele toestellen, diensten, en applicaties. Het tweede deel van dit doctoraatsproefschrift focust daarom op de analyse van de QoE in de context van mobiele videodiensten.

De eerste vier hoofdstukken van dit proefschrift behandelen het topic "persoonlijke aanbevelingen" als een hulpmiddel om onlinediensten te optimaliseren. Hoofdstuk 1 introduceert aanbevelingssystemen als onderzoeksdomein door een overzicht te geven van de verschillende technieken die beschikbaar zijn voor het genereren van persoonlijke aanbevelingen. Vervolgens zoomt dit hoofdstuk in op twee heikele punten van aanbevelingssystemen, die ook verder behandeld worden in hoofdstuk 2, 3, en 4. De evaluatie van aanbevelingen is een eerste heikel punt. Diverse kwalitatieve eigenschappen van aanbevelingssystemen, zoals nauwkeurigheid, diversiteit, dekkingsgraad, en het verrassingsaspect worden daarom besproken. Het tweede heikel punt is het genereren van groepsaanbevelingen. Dit zijn aanbevelingen die niet bedoeld zijn voor individueel gebruik maar eerder als suggesties voor een groep personen. Een typisch scenario voor het gebruik van groepsaanbevelingen is de selectie van een film samen met familie of vrienden.

Hoofdstuk 2 presenteert de resultaten van de evaluatie van 'PersonalTV', wat kan beschouwd worden als een case study van een aanbevelingssysteem. PersonalTV is een onlinevideodienst die gebruikers persoonlijke aanbevelingen aanbiedt. PersonalTV werd geëvalueerd vanuit het standpunt van de gebruiker door een panel van proefpersonen; d.w.z. een evaluatiemethode die inzichten verschaft in de interacties en ervaringen van gebruikers met deze dienst. De resultaten van deze studie tonen de consistentie van het consumptiepercentage - dit is de fractie van de video die werkelijk bekeken is - met de tevredenheid van de gebruiker betreffende het inhoudelijke aspect van de video. Met andere woorden, een video (bijna) volledig bekijken wijst op een gebruiker die tevreden is over de inhoud, terwijl het vroegtijdig onderbreken van een video overeenstemt met een minder tevreden gebruiker. Dit resultaat bevestigt de hypothese dat objectieve gebruikersinteracties (impliciete feedback) en subjectieve evaluaties (expliciete feedback) convergeren, wat impliceert dat het consumptiepercentage kan gebruikt worden als een indirect beoordelingsmechanisme. De bevindingen van deze studie en de verzamelde kwalitatieve feedback van de proefpersonen kunnen gebruikt worden voor het verbeteren en optimaliseren van de gebruikerservaring bij aanbevelingssystemen.

In hoofdstuk 3 worden de suggesties afkomstig van vijf verschillende aanbevelingsalgoritmen door gebruikers geëvalueerd in termen van nauwkeurigheid, bekendheid, nieuwigheid, diversiteit, transparantie, tevredenheid, vertrouwen, en bruikbaarheid. Deze evaluatie werd uitgevoerd in de context van een Belgische website voor culturele evenementen. Elke proefpersoon kreeg een lijst met acht aanbevelingen, gegenereerd door één van de algoritmen, en werd vervolgens gevraagd om een onlinevragenlijst in te vullen om zo de kwalitatieve aspecten van de aanbevelingen te beoordelen. De beste resultaten op alle aspecten met uitzondering van diversiteit worden behaald door het hybride algoritme, dat de aanbevelingen van het content-gebaseerde algoritme en het collaboratieve algoritme combineert. In termen van diversiteit worden de beste resultaten verkregen met een lijst van willekeurige aanbevelingen, die uiteraard bestaat uit een zeer diverse verzameling van items. Een offline evaluatie in termen van accuraatheid bevestigt in grote mate de resultaten van de gebruikersevaluatie. Daarnaast werd ook de relatie tussen de verschillende kwalitatieve aspecten van aanbevelingssystemen onderzocht. De resultaten tonen dat de nauwkeurigheid en transparantie van aanbevelingen een grote invloed hebben op de tevredenheid van de gebruiker.

In hoofdstuk 4 worden twee bestaande strategieën voor groepsaanbevelingen gecombineerd met vijf algoritmen, met als doel het genereren van groepsaanbevelingen voor verschillende samenstellingen van de groep. De eerste strategie combineert de aanbevelingen van individuele gebruikers tot aanbevelingen voor de volledige groep. De tweede strategie combineert de voorkeuren van individuele gebruikers tot een model dat de voorkeuren van de groep weerspiegelt. De groepsaanbevelingen worden geëvalueerd in termen van nauwkeurigheid, diversiteit, dekkingsgraad, en het verrassingsaspect via een offline evaluatie. De resultaten tonen dat de strategie die de meest nauwkeurige groepsaanbevelingen oplevert, afhankelijk is van het algoritme dat gebruikt wordt voor het genereren van aanbevelingen

voor individuele gebruikers. Daarom wordt in dit hoofdstuk een nieuwe strategie voor groepsaanbevelingen voorgesteld. Deze nieuwe strategie combineert de twee bestaande strategieën voor groepsaanbevelingen en presteert beter dan elk van hen in termen van nauwkeurigheid. Voorts wordt ook de invloed van de grootte en samenstelling van de groep op de kwaliteit van de aanbevelingen besproken. De nauwkeurigheid van de aanbevelingen neemt af naarmate de groep groter wordt. Voor grotere groepen wordt het immers moeilijker voor een aanbevelingssysteem om, vertrekkende van de (mogelijks tegenstrijdige) voorkeuren van de groepsleden, tot een consensus te komen. De nauwkeurigheid van de aanbevelingen neemt toe naarmate de groepsleden meer gelijkaardige voorkeuren hebben. Het vinden van aanbevelingen die alle leden van de groep kunnen bekoren is immers gemakkelijker als de groepsleden gelijkaardige interesses hebben.

De volgende vier hoofdstukken van dit doctoraatsproefschrift behandelen de analyse van de QoE, met als doel het optimaliseren van QoE parameters om zo onlinediensten te verbeteren. Hoofdstuk 5 introduceert het concept QoE en de noodzakelijke terminologie waarop verder gebouwd wordt in hoofdstuk 6, 7, en 8. Dit hoofdstuk geeft een overzicht van onderzoek rond QoE en positioneert het concept QoE ten opzichte van kwaliteitsparameters of *Quality of Service* parameters, subjectieve en objectieve kwaliteitsmetrieken, en niet-technische aspecten van een onlinedienst. Aangezien dit proefschrift focust op de QoE in de context van mobiele videodiensten, wordt een overzicht gegeven van bestaande methoden voor het evalueren van de videokwaliteit en van bestaande studies over de QoE bij mobiele diensten. Tenslotte worden beide experimentele opstellingen voor het evalueren van de QoE vergeleken, namelijk traditionele laboratorium experimenten met gecontroleerde parameters en meer waarheidsgetrouwe, zogenaamde 'living lab' experimenten.

Hoofdstuk 6 beschrijft de resultaten van een laboratorium experiment dat analyseert wanneer de kwaliteit van een mobiele videodienst onaanvaardbaar wordt voor gebruikers. De gecontroleerde omgeving van dit experiment maakte het mogelijk om bepaalde technische parameters te manipuleren, zoals de bandbreedte van de connectie die gebruikt wordt voor het zenden van de video naar een mobiel toestel. Objectieve technische parameters die opgemeten werden tijdens het experiment worden in verband gebracht met subjectieve evaluaties afkomstig van een panel van testgebruikers. Dit resulteert in een model voor het kwantificeren van de aanvaardbaarheid van onderbrekingen tijdens het bekijken van video's. Dit model biedt inzicht in de QoE door het inschatten van de kans dat gebruikers de kwaliteit van een mobiele videodienst aanvaardbaar achten, en dit als een functie van het aantal onderbrekingen tijdens het afspelen van de video.

Hoofdstuk 7 behandelt de resultaten van een verkennende studie van de QoE, meer specifiek van de subjectieve evaluaties in termen van diverse kwalitatieve aspecten met betrekking tot het bekijken van video's op een mobiel toestel in een 'living lab' context. Zeker wanneer de focus ligt op mobiele applicaties en diensten kan onderzoek in een dergelijke waarheidsgetrouwe omgeving een interessante aanvulling zijn voor onderzoek in laboratoria. Op basis van objectieve technische parameters en subjectieve evaluaties wordt de videokwaliteit, zoals waargenomen

door de gebruiker, onderzocht. Dit resulteert in een model voor het inschatten van de subjectieve gebruikersperceptie van de laadsnelheid en een model voor het inschatten van de subjectieve gebruikersperceptie van visuele storing tijdens het mobiel bekijken van video's. Tot slot presenteert dit hoofdstuk een beslissingsboom voor het kwantificeren van de QoE tijdens het mobiel bekijken van video's op basis van technische parameters van de videosessie en het netwerk. Deze resultaten verschaffen applicatieontwerpers en onlinedienstverleners een leidraad die verduidelijkt welke en hoe technische parameters de QoE beïnvloeden, maar ook hoe die parameters moeten aangepast worden om de QoE te optimaliseren.

Hoofdstuk 8 maakt de brug tussen het eerste deel van dit doctoraatsproefschrift omtrent aanbevelingssystemen en het tweede deel, dat de analyse van de QoE bespreekt. Meer concreet, dit hoofdstuk behandelt de invloed van de QoE op de expliciete feedback van de gebruikers, die gehanteerd wordt bij het genereren van aanbevelingen. De resultaten tonen aan dat gebruikers geneigd zijn om een video inhoudelijk een lagere waardering te geven (expliciete feedback) indien de kwaliteit tijdens het afspelen van de video niet optimaal is. Dus de expliciete feedback van een gebruiker weerspiegelt de voorkeuren van die gebruiker niet helemaal correct. Daarom wordt in dit hoofdstuk een model voorgesteld om de invloed van een wisselende QoE op de expliciete feedback van de gebruikers te corrigeren. Dit model kan in aanbevelingssystemen gebruikt worden om de nauwkeurigheid van de expliciete feedback te verbeteren en als gevolg ook de nauwkeurigheid van de aanbevelingen.

Tot slot worden in hoofdstuk 9 de algemene conclusies van dit doctoraatsproefschrift gepresenteerd, en worden opportuniteiten voor toekomstig onderzoek kort beschreven.

English summary

Over the last decade, the ubiquity of the Internet in everyday life has stimulated the growth of a wide variety of online service, such as online shops, informative web sites, video delivery services, etc. As a result, users are confronted with the problem of information overload: although an abundance of information is available, obtaining useful and relevant information is often a difficult task. This problem can be addressed by recommender systems, which assist users in discovering and selecting the most interesting information or items. Recommender systems offer users personalized suggestions based on their preferences, and optimize in this way the usability and usefulness of online services. The first part of this dissertation focusses on recommender systems an their evaluation in terms of different qualitative aspects.

However, the content (i.e., the available items or information) and the features of an online service are no longer the only differentiator. The experience that the user has while using an online service has become an important factor in the design, development, and optimization process of a service. Nevertheless, quantifying the Quality of Experience (QoE) remains challenging, especially in the mobile media domain, which is characterized by an exponential growth in the number of mobile devices, services, and applications, and by the availability of various new technologies and access networks. The second part of this dissertation focusses on the analysis of the QoE of video delivery services in a mobile environment.

The first four chapters of this dissertation cover the topic of personalized recommendations as a tool to optimize online services. Chapter 1 introduces the research area of recommender systems by providing an overview of the different techniques that are available to generate personalized recommendations. Subsequently this chapter zooms in on two issues in the domain recommender systems that are tackled in Chapter 2, 3, and 4. The first issue is the evaluation of recommendations. Various qualitative aspects of recommender systems, such as accuracy, diversity, coverage, and serendipity, are discussed. The second issue is the generation of group recommendations, i.e., recommendations that are not intended for individual usage but rather for consumption in group. A typical use case for group recommendations is the selection of a movie with friends or family.

Chapter 2 presents the results of the evaluation of the 'PersonalTV' service, a recommender system case study. PersonalTV is an online video delivery service, offering personalized recommendations to its users. The PersonalTV service was evaluated from a user point of view by a panel of test subjects, thereby providing insights into the users' interaction behaviour and experience with the system. The

results of this study demonstrate the consistency between the consumption percentage, i.e., the fraction of the video that is actually watched, and the user's satisfaction with the content: (nearly) complete viewing of a video corresponds to a significantly higher satisfaction with the content than incomplete viewing, thereby implying that consumption percentage can be used as an indirect rating mechanism. This confirms the hypothesis of convergence of measured objective user interaction (implicit feedback) and subjective evaluations (explicit feedback). The results of this study and the gathered qualitative user feedback can serve as input for improving and optimizing users' experiences with recommender systems.

Chapter 3 discusses a user-centric evaluation of five different recommendation algorithms in terms of accuracy, familiarity, novelty, diversity, transparency, satisfaction, trust, and usefulness. For this evaluation, which was performed in the context of a Belgian cultural event website, each of the test subjects received a list of eight recommendations generated by one of the algorithms. Then, the test subjects were asked to fill in an online questionnaire that addressed the qualitative aspects of their recommendation list. The results clearly show that the hybrid algorithm, which combines the recommendations of the content-based and collaborative filtering algorithm, outperforms every other algorithm except for the diversity aspect. In terms of diversity, the random recommendations turn out best, which are of course a very diverse set of items. An offline evaluation in terms of accuracy confirms the results of the user evaluation to a large extend. In addition, the relationships between the different qualitative aspects of recommender systems are investigated. The results indicate that the accuracy and transparency are influential predictors of the user satisfaction.

In Chapter 4, two existing strategies for generating group recommendations are combined with five algorithms in order to calculate group recommendations for different group compositions. The first strategy aggregates the users' individual recommendations into recommendations for the whole group (aggregating recommendations). The second strategy aggregates the users' individual preferences into a preference model of the group (aggregating preferences). The group recommendations are evaluated in terms of accuracy, diversity, coverage, and serendipity by means of an offline evaluation. The results show that the aggregation strategy that produces the most accurate group recommendations depends on the algorithm that is used for generating individual recommendations. For that reason, this chapter proposes a new aggregation strategy, combining the two strategies and outperforming each individual strategy in terms of accuracy. In addition, the influence of the size and composition of the group on the quality of the recommendations is discussed. The accuracy of the group recommendations decreases as the group size increases, since mediating the potentially contrasting preferences of the group members becomes more difficult for larger groups. The accuracy of the group recommendations increases as the similarity between members of the group increases, because finding group recommendations that satisfy all group members is easier as the group members are more similar to each other.

The next four chapters of this dissertation cover the topic of QoE analysis as a means for optimizing QoE parameters and thereby improving online services.

Chapter 5 introduces the concept of QoE and the necessary terminology on which is built in Chapter 6, 7, and 8. This chapter outlines the research regarding QoE and positions the concept QoE with respect to Quality of Service (QoS) parameters, subjective and objective quality metrics, and non-technical aspects. Since this dissertation focusses on the QoE during video watching in a mobile environment, an overview of existing video quality assessment methods is provided and existing studies covering QoE in a mobile environment are reviewed. Lastly, both experimental settings to evaluate QoE, traditional test beds with controlled laboratory parameters and real-life (so called 'living lab') experiments in the field, are compared.

Chapter 6 describes the results of a controlled laboratory experiment that explores the thresholds at which the technical quality of a mobile video service becomes unacceptable for users. The controlled environment of this experiment allowed to manipulate the bandwidth of the data connection used to transfer the videos to the mobile device. Objective technical parameters measured during the experiment are combined with subjective evaluations by a user panel, resulting in a model for quantifying the acceptability of video interruptions. This model provides insights into the QoE by estimating the probability that users will accept the quality of a mobile video session as a function of the number of rebuffering interruptions during video playback.

Chapter 7 presents results from an exploratory study on QoE, and more specifically, on the subjective evaluations of various quality aspects during video watching on a mobile device, in a living lab context. Especially when focusing on mobile applications and services, research in realistic settings may complement controlled laboratory testing. The perceived quality of a mobile video session is investigated based on measurements of technical parameters and subjective evaluations, resulting in a model for estimating the user's subjective evaluation of the loading speed and a model for estimating the subjectively-perceived distortion during mobile video watching. Finally, this chapter presents a decision tree for quantifying QoE during mobile video watching based on technical parameters of the video session and the network. These results provide application developers and service providers a tool that clarifies which and how technical parameters influence the QoE and how the parameters have to be adapted to optimize the QoE.

Chapter 8 makes the bridge between the first part of this dissertation concerning recommender systems, and the second part, that is about QoE. More concretely, this chapter discusses the influence of QoE on the user's explicit feedback that is used for generating recommendations. The results show that users tend to give a lower evaluation for the content (explicit feedback) if the quality of the video playback is not optimal. This means that the user's explicit feedback does not correctly reflect the preferences of the user, thereby possibly effecting the accuracy of recommender systems. Therefore, this chapter proposes a model, which can be used in recommender systems, to correct the user's explicit feedback for video content by considering the influence of a varying QoE.

Finally in Chapter 9, overall conclusions are presented and opportunities for future research are briefly described.

Publications

A1

(publications in journals listed in the ISI Web of Science)

As first author

[1] **T. De Pessemier**, T. Deryckere, K. Vanhecke, and L. Martens, *Proposed Architecture and Algorithm for Personalized Advertising on iDTV and Mobile Devices.* Published in IEEE Transactions on Consumer Electronics, 54(2):709-713, May 2008.

[2] **T. De Pessemier**, K. De Moor, I. Ketykó, W. Joseph, L. De Marez, and L. Martens, *Investigating the Influence of QoS on Personal Evaluation Behaviour in a Mobile Context.* Published in Springer Multimedia Tools and Applications, 57(2):335-358, March 2012.

[3] **T. De Pessemier**, S. Coppens, K. Geebelen, C Vleugels, S. Bannier, E. Mannens, K. Vanhecke, and L. Martens, *Collaborative Recommendations with Content-based Filters for Cultural Activities via a Scalable Event Distribution Platform.* Published in Springer Multimedia Tools and Applications, 58(1):167-213, May 2012.

[4] **T. De Pessemier**, K. De Moor, W. Joseph, L. De Marez, and L. Martens, *Quantifying Subjective Quality Evaluations for Mobile Video Watching in a Semi-Living Lab Context.* Published in IEEE Transactions on Broadcasting, 58(4):580-589, December 2012.

[5] **T. De Pessemier**, K. De Moor, W. Joseph, L. De Marez, and L. Martens, *Quantifying the Influence of Rebuffering Interruptions on the User's Quality of Experience During Mobile Video Watching.* Published in IEEE Transactions on Broadcasting, 59(1):47-61, March 2013.

[6] **T. De Pessemier**, S. Dooms, and L. Martens, *Comparison of group recommendation algorithms.* Accepted for publication in Springer Multimedia Tools and Applications, 2013.

As co-author

[1] S. Coppens, E. Mannens, **T. De Pessemier**, K. Geebelen, H. Dacquin, D. Van Deursen, and R. Van de Walle, *Unifying and Targeting Cultural Activities via Events Modelling and Profiling.* Published in Springer Multimedia Tools and Applications, 57(1):199-236, March 2012.

[2] E. Mannens, S. Coppens, **T. De Pessemier**, H. Dacquin, D. Van Deursen, R. De Sutter, and R. Van de Walle, *Automatic News Recommendations via Aggregated Profiling.* Published in Springer Multimedia Tools and Applications, 63(2):407-425, March 2013.

[3] S. Dooms, **T. De Pessemier**, D. Verslype, J. Nelis, J. De Meulenaere, W. Van den Broeck, L. Martens, and C. Develder, *OMUS: an Optimized Multimedia Service for the Home Environment.* Accepted for publication in Springer Multimedia Tools and Applications, 2013.

A2

(publications in journals not listed in the ISI Web of Science)

As co-author

[1] K. De Moor, **T. De Pessemier**, P. Mechant, C. Courtois, A. Juan, L. De Marez, and L. Martens, *Users' (Dis)satisfaction with the PersonalTV Application: Combining Objective and Subjective Data.* Published in ACM Computers in Entertainment Special Issue EuroITV'10, 9(3): 18:1-18:22, November 2011.

P1

(publications in international conferences listed in the ISI Web of Science)

As first author

[1] **T. De Pessemier**, T. Deryckere, and L. Martens, *Context Aware Recommendations for User-Generated Content on a Social Network Site,* In Proceedings of the seventh European Interactive Television Conference, EuroITV '09, pages 133-136, Leuven, Belgium, June 2009. ACM.

[2] **T. De Pessemier**, K. Vanhecke, S. Dooms, T. Deryckere, and L. Martens, chapter *Extending User Profiles in Collaborative Filtering Algorithms to Alleviate the Sparsity Problem*, in Web Information Systems and Technologies, edited by J. Filipe, and J. Cordeiro, Vol. 75, pages 230-244, 2011, Springer Berlin Heidelberg.

C1

(publications in other international conferences)

As first author

[1] **T. De Pessemier**, M. Ide, T. Deryckere, and L. Martens, *Consumption context and personalization,* In Poster Proceedings of the sixth European Interactive Television Conference, EuroITV '08, pages 1-4, Salzburg, Austria, July 2008. ACM.

[2] **T. De Pessemier**, K. Vanhecke, S. Dooms, T . Deryckere, and L. Martens, *Probability-Based Extended Profile Filtering, An Advanced Collaborative Filtering Algorithm for User-Generated Content,* In Proceedings of the 6th International Conference on Web Information Systems and Technologies, WEBIST 2010, pages 219-226, Valencia, Spain, April 2010. INSTICC.

[3] **T. De Pessemier**, and L. Martens, *Extending the Bayesian Classifier to a Context-Aware Recommender System for Mobile Devices,* In Proceedings of the fifth International Conference on Internet and Web Applications and Services, ICIW 2010, pages 242-247, Barcelona, Spain, May 2010. IEEE.

[4] **T. De Pessemier**, S. Dooms, T. Deryckere and L. Martens, *Time Dependency of Data Quality for Collaborative Filtering Algorithms,* In Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys '10, pages 281-284, Barcelona, Spain, September 2010. ACM.

[5] **T. De Pessemier**, S. Coppens, E. Mannens, S. Dooms, K. Geebelen, and L. Martens, *An Event Distribution Platform for Recommending Cultural Activities,* In Proceedings of the 7th International Conference on Web Information Systems and Technologies, WEBIST 2011, pages 231-236, Noordwijkerhout, Netherlands, May 2011. INSTICC.

[6] **T. De Pessemier**, K. Vanhecke, S. Dooms, and L. Martens, *Content-Based Recommendation Algorithms on the Hadoop Map-Reduce Framework,* In Proceedings of the 7th International Conference on Web Information Systems and Technologies, WEBIST 2011, pages 237-240, Noordwijkerhout, Netherlands, May 2011. INSTICC.

[7] **T. De Pessemier**, K. De Moor, A. J. Verdejo, D. Van Deursen, W. Joseph, L. De Marez, and L. Martens, *Quantifying QoE of Mobile Video Consumption in a Real-Life Setting Drawing on Objective and Subjective Parameters,* The International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB 2011, pages 1-6, Nürnberg, Germany, June 2011. IEEE.

[8] **T. De Pessemier**, S. Dooms, J. Roelandts, and L. Martens, *Analysis of the Information Value of User Connections for Video Recommendations in a Social Network*, In Proceedings of the 2nd International Workshop on Future Television held at the 9th European Conference on Interactive TV and Video (FutureTV-2011), EuroITV '11, pages 1-7, Lisbon, Portugal, June 2011. ACM.

[9] **T. De Pessemier**, K. De Moor, A. J. Verdejo, D. Van Deursen, W. Joseph, L. De Marez, and L. Martens, *Exploring the Acceptability of the Audiovisual Quality for a Mobile Video Session Based on Objectively Measured Parameters,* In Proceedings of the third International Workshop on Quality of Multimedia Experience, QoMEX 2011, pages 125-130, Mechelen, Belgium, September 2011. IEEE.

[10] **T. De Pessemier**, L. Van Acker, E. Van Dijck, K. Slegers, W. Joseph, and L. Martens, *A Mobile Conversation Assistant to Enhance Communications for Hearing-impaired Children,* In Proceedings of the 8th International Conference on Web Information Systems and Technologies, WEBIST 2012, pages 775-780, Porto, Portugal, April 2012. INSTICC.

[11] **T. De Pessemier**, S. Dooms, and L. Martens, *Design and Evaluation of a Group Recommender System*. In Proceedings of the sixth ACM conference on Recommender Systems, RecSys '12, pages 225-228, Dublin, Ireland, September 2012. ACM.

[12] **T. De Pessemier**, K. De Moor, L. De Marez, L. Martens, and W. Joseph, *Quantifying QoE Indicators in a Living Lab Context,* 2nd IEEE BTS GOLD Workshop: Next Generation Broadcasting, pages 1-3, Cagliari, Italy, March 2013. IEEE.

[13] **T. De Pessemier**, S. Dooms, K. Vanhecke, B. Matté, E. Meyns, and L. Martens, *Context-Aware Recommendations through Activity Recognition,* In Proceedings of the 9th International Conference on Web Information Systems and Technologies, WEBIST 2013, pages 481-490, Aachen, Germany, May 2013. INSTICC.

[14] **T. De Pessemier**, K. De Moor, L. De Marez, L. Martens, and W. Joseph, *Modeling Subjective Quality Evaluations for Mobile Video Watching in a Living Lab Context,* The International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB 2013, pages 1-6, Uxbridge, West London, UK, June 2013. IEEE.

As co-author

[1] K. Berte, K. Vanhecke, J. Pelssers, W. Holvoet, **T. De Pessemier**, G. Jans, V. Verbrugghe, T. Deryckere, P. Leroux, L. Martens, F. De Turck, P. Demeester, and E. De Bens, *Advertising in a Digital Media Environment (ADME): An Interdisciplinary Approach to a User-centered Advertising Model for IDTV,* Budapest workshop 2008: Digital television revisited: linking users, markets and policies, COST Action 298 'Participation in the Broadband Society', pages 123-131, Budapest, Hungary, May 2008.

[2] E. Mannens, S. Coppens, **T. De Pessemier**, K. Geebelen, H. Dacquin, and R. Van de Walle, *Unifying and Targeting Cultural Activities via Events Modelling and Profiling,* In Proceedings of the 1st ACM international workshop on Events in multimedia, EiMM '09, pages 33-40, Beijing, China, October 2009. ACM.

[3] K. De Moor, **T. De Pessemier**, P. Mechant, C. Courtois, A. Juan, L. Demarez, and L. Martens, *Evaluating a Recommendation Application for Online Video Content: an Interdisciplinary Study*, In Proceedings of the eighth European Interactive Television Conference, EuroITV '10, pages 115-122, Tampere, Finland, June 2010. ACM.

[4] A. J. Verdejo, K. De Moor, Ketykó, K. T. Nielsen, J. Vanattenhoven, **T. De Pessemier**, W. Joseph, L. Martens, L. De Marez, *QoE Estimation of a Location-Based Mobile Game using On-Body Sensors and QoS-related Data,* 2010 IFIP Wireless Days conference - Wireless Multimedia and Entertainment, WD 2010, pages 1-5, Venice, Italy, October 2010. IEEE.

[5] I. Ketykó, K. De Moor, **T. De Pessemier**, A. J. Verdejo, K. Vanhecke, W. Joseph, L. Martens, and L. De Marez, *QoE Measurement of Mobile YouTube Video Streaming,* In Proceedings of the 3rd workshop on Mobile video delivery, MoViD '10, pages 27–32, Firenze, Italy, October 2010. ACM.

[6] E. Mannens, S. Coppens, **T. De Pessemier**, H. Dacquin, D. Vandeursen, and R. Van de Walle, *Automatic News Recommendations via Profiling,* In Proceedings of the 3rd International Workshop on Automated Information Extraction in Media Production, AIEMPro '10, pages 45-50, Florence, Italy, October 2010. ACM.

[7] S. Dooms, **T. De Pessemier**, and L. Martens, *An Online Evaluation Of Explicit Feedback Mechanisms for Recommender Systems,* In Proceedings of the 7th International Conference on Web Information Systems and Technologies, WEBIST 2011, pages 391-394, Noordwijkerhout, Netherlands, May 2011. INSTICC.

[8] S. Dooms, **T. De Pessemier**, and L. Martens, *A File-Based Approach for Recommender Systems in High-Performance Computing Environments,* In Proceedings of the 22nd International Conference on Database and Expert Systems Applications (RSMEETDB 2011 - Workshop on Recommender Systems meet Databases), DEXA 2011, pages 529-533, Toulouse, France, August 2011. IEEE.

[9] S. Dooms, **T. De Pessemier**, and L. Martens, *A User-centric Evaluation of Recommender Algorithms for an Event Recommendation System,* In Proceedings of the RecSys 2011 Workshop on Human Decision Making in Recommender Systems (Decisions@RecSys'11) and User-Centric Evaluation of Recommender Systems and Their Interfaces - 2 (UCERSTI 2), RecSys '12, pages 67-73, Chicago, IL, October 2011. ACM.

[10] S. Dooms, **T. De Pessemier**, and L. Martens, *Caching Strategies for In-memory Neighborhood-based Recommender Systems,* In Proceedings of the 9th International Conference on Web Information Systems and Technologies, WEBIST 2013, pages 435-440, Aachen, Germany, May 2013. INSTICC.

(publications in national conferences)

As first author

[1] **T. De Pessemier**, and L. Martens, *A Profile-Based Recommendation System for TV-Anytime Annotated Content,* 8th UGent - Firw PhD symposium, Ghent, Belgium, December 2007.

[2] **T. De Pessemier**, *Reducing the Complexity of the Content Selection Process by Recommendation Techniques, Social Networks and Consumption Contexts*, IBBT Friday Food event, Ghent, Belgium, March 2009.

[3] **T. De Pessemier**, *Recommender Engine Ins en Outs*, Cupid Salon, Brussels, Belgium, June 2010.

[4] **T. De Pessemier**, and L. Martens, *Data Analysis for Collaborative Filtering Systems,* 11th UGent - Firw PhD symposium, Ghent, Belgium, December 2010.

As co-author

[1] K. De Moor, **T. De Pessemier**, P. Mechant, and C. Courtois, *Harnessing Implicit and Explicit User Feedback for the Evaluation of a Facebook Application,* Etmaal van de Communicatiewetenschap, Ghent, Belgium, Februari 2010.

[2] S. Dooms, **T. De Pessemier**, and L. Martens, *Demonstrating Contextual Group Recommendations for Media in a Home Environment,* In Proceedings of the 12th Dutch-Belgian Information Retrieval Workshop, DIR2012, pages 83-84, Ghent, Belgium, February 2012.

Awards

[1] Winner of the best paper award for
K. De Moor, **T. De Pessemier**, P. Mechant, C. Courtois, A. Juan, L. Demarez, and L. Martens *Evaluating a Recommendation Application for Online Video Content: an Interdisciplinary Study*, In Proceedings of the eighth European Interactive Television Conference, EuroITV '10, pages 115-122, Tampere, Finland, June 2010. ACM.

Part I

Recommender Systems

1

Introduction to recommender systems

1.1 Recommender systems

Recommender systems are software tools and techniques providing suggestions for items to be of interest to a user [1]. "Item" is the general term used to denote what the system recommends to users. The suggestions provided are aimed at supporting their users in various decision-making processes, such as what items to buy, what movies to watch, or what news to read. These suggestions are often offered as a ranked list of items. In performing this ranking, recommender systems try to predict what the most suitable items are, based on the user's preferences. These user preferences are gathered by collecting users' explicit ratings for items or by interpreting user actions such as clicks or purchases [2]. A rating or a user interaction with an item is commonly referred to as a "consumption" of an item.

Recommender systems are a relatively new research domain. Joseph A. Konstan[1] stated: *"It was a seductively simple idea that emerged in the early 1990s - to harness the opinions of millions of people online in an effort to help all of us find more useful and interesting content"* [3]. The idea and terminology of Collaborative Filtering (CF) was born [4], and this was the beginning of an increasingly growing research activity in the domain of personalization and recommender systems.

[1] Joseph A. Konstan is a professor at the University of Minnesota and a member of the GroupLens research lab. He is considered an authority in the field of human-computer interaction and recommender systems.

In 1994, the GroupLens system, named after the research lab, demonstrated the possibilities of CF in a network in the context of Usenet news messages [5]. In 1997, the GroupLens lab launched MovieLens[2], a web-based recommender system for movies that was built for research purposes [6]. MovieLens has several thousand regular users, who rate movies and in return receive personal recommendations for other movies. The GroupLens lab has collected and made available rating data sets from the MovieLens web site. From then on, these data sets are used as a common tool by numerous researchers to test and evaluate recommender systems. Based on the research findings of the MovieLens system [6], several critical issues in the domain of recommender systems have been identified such as the evaluation of recommender systems and the variety of evaluation metrics (as discussed in Section 1.2), and the use of implicit feedback. Implicit feedback are observational measures of the user's preference for an item, e.g., the time spent on reading an article, watching a movie, or listening to a song. If implicit feedback is consistent with explicit feedback (such as explicit ratings), it can be used as an alternative or complementary knowledge source (as discussed in Chapter 2). Through PolyLens [7], an extension of MovieLens, the issue of group recommendations is tackled. Group recommendations are used to suggest items for groups of users (e.g., a family), rather than for individuals (as discussed in Section 1.3).

A further boost in the research and development of recommender systems came from the Netflix prize competition, which began in October 2006. Netflix, an American provider of on-demand Internet streaming media, organised the Netflix Prize[3] to find a better algorithm to predict user preferences and beat its existing Netflix movie recommendation algorithm. The Netflix Prize was an open competition with a grand prize of $ 1,000,000 attracting thousands of scientists, students, and engineers in the field of recommender systems. This competition was important for the evolution of recommender systems for two reasons: (1) for the first time, the research community gained access to a large-scale data set containing 100 million movie ratings; (2) all algorithms were judged on common data using the same evaluation procedure and metrics, providing a common ground to compare the efficiency of the different techniques [2].

Over the years, several different approaches for generating recommendations have been proposed. These can be classified according to a taxonomy that was based on the work of Burke [8] and that has become a classical way of distinguishing between recommender systems and referring to them [2]. Six different classes of recommendation techniques can be distinguished based on the knowledge source[4].

[2] http://www.movielens.umn.edu

[3] http://www.netflixprize.com/

[4] There is another knowledge source net yet included in the classification: context, which is becoming important, particularly for mobile applications.

1.1.1 Demographic recommendations

This is a simple recommendation technique that suggests items based on the demographic profile of the user. These demographic data can include the user's gender, age, home town and country, language, etc. Different demographic groups receive different recommendations. Many websites use this simple solution to offer a "personalized" content offer. E.g., users are redirected to a particular website based on their language or country. These approaches have been quite popular in the marketing literature, but have received little attention in the field of recommendation algorithms [9].

1.1.2 Knowledge-based recommendations

Knowledge-based systems generate recommendations using specific domain knowledge about how certain item features meet user preferences, and ultimately, how the item is of interest to the user.

One type of knowledge-based systems is based on *case-based reasoning* [10, 11]. Cased-based reasoning focusses on the reuse of expertise, which is modelled as cases. Aamodt and Plaza refer to case-based reasoning as a problem-solving paradigm that uses the specific knowledge gathered by solving concrete problem situations [12]. In practice, a case-based recommender system estimates how much the user's needs or preferences (problem description) match the potential recommendations (solutions of the problem) based on previous consumption behaviour (previous cases).

Constraint-based recommenders are another type of knowledge-based systems [13]. Constraint-based recommenders exploit predefined knowledge bases that contain explicit rules about how to relate customer requirements with item features [2]. E.g., a user might be interested to buy products with a certain set of features and within a specific price range.

Knowledge-based systems typically work better than other types of recommenders if limited data is available, i.e., if the system cannot rely on the existence of a user history. But if the knowledge-based system is not designed to learn from ratings or user actions, other, more intelligent recommendation algorithms (such as CF) will soon outperform the knowledge-based system.

1.1.3 Community-based recommendations

People tend to rely more on recommendations from their friends than on recommendations from similar but anonymous individuals [14]. Moreover, information about the user and the user's friends can be gathered from popular social networks. This knowledge is used in community-based systems, or as they are often called, social recommender systems [15]. So, community-based recommender systems

generate recommendations based on the user's connections or relations with other users or friends in the social network, the preferences of these friends, and sometimes also on a value for the user's trust in each of his/her friends. Research in the domain of community-based recommenders is still in an early phase and results about the system performance are dependent on the specific case [2].

1.1.4 Content-based recommendations

The concept of a content-based recommender system is to suggest the items that are similar to the items that the user liked in the past. The recommender system learns which (type of) items the user likes based on the user's consumption behaviour and the attributes that describe an item. These attributes provide useful information about the item, such as the title, a description, keywords, categories, etc., and can be denoted as metadata. The items' attributes are used to construct a user profile, in which the personal preferences and interests of the user are stored. Then, the main operation performed by a content-based recommender consists in matching the attributes of the user's profile with the attributes of (unexplored) content items, with the aim of finding interesting recommendations that match the user's preferences.

Different techniques are used for content-based recommender systems: probabilistic models based on the naïve Bayes assumption [16], Support Vector Machines (SVM) [17], Rocchio's relevance feedback method [18], and nearest neighbour methods in the vector space model [19].

Based on the naïve Bayes assumption, a probabilistic model can be generated to estimate the probability that an item belongs to a certain class (e.g., the class of relevant items) [2]. The standard SVM is based on the concept of decision planes that define decision boundaries. A decision plane separates the item space in two different classes (e.g., relevant and not relevant). The SVM takes a set of input data (attributes of the item) and predicts for each item at which side of the decision plane it belongs. Each side of the decision plane corresponds to an item class specifying the relevance of the item [20]. Rocchio's relevance feedback method helps users to incrementally refine their queries by giving feedback on whether the retrieved items are relevant [3]. The vector space model is a spatial representation of items, in which each item is represented by a n-dimensional vector. The similarity of two items in the vector space model is determined by means of a similarity measure, such as the cosine similarity. Recommendations can be derived by calculating the similarity between an item and the user profile, which is also represented by a vector in the space model [2].

The big advantage of content-based systems is that they do not require a large community of users to achieve a reasonable performance. In addition, new items can be immediately recommended once the item attributes are available. In other words, content-based recommendations do not suffer from the *first-rater* or *new-*

item problem. Content-based recommendations are also easy to explain based on the user's historical consumption behaviour, thereby improving the "transparency" of the system. Transparency determines whether or not a system allows users to understand its inner logic, i.e., why a particular item is recommended to them [21].

A disadvantage of these content-based algorithms is the dependence on the availability of (manually created or automatically extracted) item metadata. For existing (commercial) services in which a recommendation system was not of primary importance in the development phase, the availability of item metadata might be a problem. In addition, domain knowledge can also be required, e.g., for books, the system needs to know that the author is an important characteristic of the item. Content-based recommenders suggest items that best match against the user profile, which is built on previous consumptions. As a result, users can receive suggestions for items that are extremely similar to the ones they just consumed. This is the problem of *over-specialization*. Finally, content-based recommenders have difficulties to generate effective recommendations for a user if only a few consumptions of that user are available. This is called the *new-user problem*. In order to really understand the preferences of the user, enough consumptions of that user have to be collected (or an initial explicit profile has to be available).

1.1.5 Collaborative recommendations

The hypothesis of systems based on CF is that if users shared the same or similar preferences in the past, they will also have similar interests in the future. The similarity in preferences of two users is calculated based on the similarity in the consumption history of the users. So, if for example, user A and B have expressed similar preferences to the same items in the past, and user A had recently consumed and liked a new item that B has not yet explored, the basic rational is to suggest this new item also to B.

CF systems have been extensively studied over the past fifteen years and today they are also in wide use in commercial services or applications [3]. Different techniques have been proposed to implement and optimize CF. Nearest Neighbourhood (NN) methods are very popular because of their simplicity, efficiency, and ability to produce accurate recommendations. In NN approaches, the user-item consumptions stored in the system are directly used to predict the user's preferences for new items. This can be done in two ways known as user-based or item-based recommendation [2].

User-based NN approaches [22], referred to as User-Based Collaborative Filtering (UBCF), predict the user's interest for an item using the ratings for this item by other users, called neighbours. These neighbours are users who have similar preferences, or in other words, users that have a similar consumption behaviour. Similarities are calculated by a similarity measure such as the Pearson correlation or the cosine similarity [3]. Item-based NN approaches [23] on the other hand, pre-

dict the user's interest for an item based on the user's ratings for similar items. In case of Item-Based Collaborative Filtering (IBCF), two items are similar if several users have consumed or evaluated these items in a similar way.

Model-based approaches use consumptions to learn a predictive model. The general idea is to model user-item interactions with factors representing latent characteristics of the users and items in the system, such as the preference class of users and the category class of items. The available consumption data is first used to train the model. Then, the model can be used to predict the user's preference for an item [2]. Various model-based approaches of CF exist, such as SVM [24] and Singular Value Decomposition (SVD) [25]. SVD is a matrix factorisation method that transforms both users and items to the same latent factor space. Subsequently, this latent factor space is used to model the consumptions by characterizing the corresponding users and items in terms of the latent factors, which are automatically inferred [2].

The big advantage of CF approaches is that they do not require any knowledge about the items (metadata) or the domain. Since recommendations are generated based on the consumption behaviour of other, similar users in the system, CF has the potential to generate less obvious and more surprising suggestions.

But CF systems also have some disadvantages. They require a large enough community of users and sufficient consumption behaviour to be able to find similar users or similar items. In real-world recommender systems, users provide ratings for only a small fraction of the potentially large amount of catalog items. A big challenge for CF systems is to generate suitable recommendations when there are relatively few ratings available, this is called the *sparsity problem*. A special case of this sparsity problem is the cold start problem [26], which focusses on the difficulty to generate recommendations for new users that have not yet consumed any item (*new-user problem*) and the issue of dealing with items that have not been consumed yet (*new-item problem*).

1.1.6 Hybrid recommendations

Different recommendation techniques have different advantages and induce different drawbacks and problems. A solution might be to combine the different techniques to eliminate these drawbacks, thereby obtaining better recommendations. E.g., if a system has a large community of users and detailed metadata about the items are available, the recommender system can be enhanced by hybridizing collaborating filtering with content-based techniques. In particular, such a hybrid recommender can overcome the new-item problem of CF by relying on the analysis of the item attributes, while taking advantage of the community knowledge. Several methods have been proposed for combining recommendation techniques in order to create a new hybrid system [8].

In research concerning recommender systems, collaborative, content-based, and hybrid recommendations received most attention because of their potential, and widespread use. Also in this dissertation, the focus is on these three types of recommendation algorithms.

1.2 Evaluating recommendations

To evaluate a recommendation algorithm, or a recommender systems as a whole, two different approaches are possible: online and offline evaluations. In an online evaluation, the recommender system is tested by real users using the real application in order to evaluate one configuration or to compare multiple alternative configurations of the system [2]. Online evaluations have several advantages. Since this approach involves real users, it delivers very trustworthy results. Moreover, it measures the performance on the real application in a real usage context. But online evaluations also induce problems and difficulties. First and foremost, a large enough user set is required to obtain relevant results; and with a limited user population, comparing many alternatives takes a long time. Secondly, since real users are involved in the experiment, testing has an impact on their experience with the service, which might be undesired. And thirdly, because these real users have to use the real application, online evaluations are not applicable to applications or services before they are launched.

More detailed results and qualitative feedback can be obtained by extending the online evaluations with a user study. Via a user study, users can be questioned before, during, or after service usage about (their experience with) the recommendation service or application. These users, also called test subjects, can provide additional relevant information about their behaviour, intentions, or expectations through self-report methods (such as questionnaires, diaries, focus groups, etc.). Although such methods can help to discover important insights, a common criticism deals with their possible subjective bias [27]. This possible bias is mainly due to the fact that self-report methods are largely based on introspection[5] and recall in memory[6]. Another concern is that, depending on the level of obtrusiveness, self-report measures might interfere or even interrupt the user's experience with a given application. Moreover, gathering test subjects that truly represent the system's real users can be difficult. And since these test subjects are usually getting paid to test the system and answer questions, user studies can also become quite expensive. Chapter 2 of this dissertation presents the results of a user study providing insights into the users' interaction behaviour with recommender systems and aiming to understand how a recommender system is experienced and evaluated from a user point of view. In Chapter 3, an online evaluation of multiple recom-

[5] Introspection is the examination of one's own conscious thoughts and feelings.
[6] Recall in memory refers to the retrieval of events or information from the past.

mendation algorithms is discussed in the context of a cultural event website. In this experiment, real users of the website were asked to evaluate the algorithms by filling in a questionnaire.

Since online evaluations with real users are mostly expensive and risky, offline evaluations are often used as a methodology for cheap and rapid experimentations [2]. These offline evaluations are typically using data sets of historical user behaviour to evaluate new recommendation algorithms or algorithm parameters. The data set is split into two disjoint sets: the training set and the test set. The training set represents the interactions that the users have already performed, and is used as input for the recommender. The test set stands for interactions that the users have not yet performed; these are unknown for the recommender and have to be predicted. The big advantage of offline evaluations is that many tests can be performed in a short time period, cheaply, and without the need of real users. However, offline evaluations have also serious disadvantages. Firstly, it is difficult to guarantee that the evaluation methodology truly captures real user behaviour. In reality (and in online evaluations), recommendations affect user behaviour; but offline evaluations merely predict user behaviour without the influence of recommendations. As a result, it is difficult to estimate the true change in user behaviour provoked by the recommendations. Secondly, the evaluation metrics and methodology used to evaluate the recommendation algorithm are not (yet) standardized, but have a serious influence on the results. Different evaluation metrics or different evaluation methodologies can lead to totally contrasting conclusions about the quality of recommendations [28]. And thirdly, not all qualitative aspects of a recommender system can be captured by an offline evaluation. E.g., the trust that users put in the recommender system cannot be evaluated by using an offline evaluation. In Chapter 4 of this dissertation, an offline evaluation of recommendation algorithms is discussed in the context of a group recommender system for audiovisual content in the home environment.

To evaluate the quality of a recommender system, different qualitative attributes can be measured. In case of an online evaluation, various qualitative attributes can be assessed by means of a subjective evaluation through a questionnaire or an interview. In case of an offline evaluation, evaluation metrics have been developed to estimate the quality of different aspects of the recommender system.

1.2.1 Accuracy

Most research regarding recommendation algorithms has focused on improving the (prediction) accuracy of recommender systems. The accuracy reflects the extent to which the recommendations match the true preferences of the user. In an offline evaluation, the accuracy attribute measures how precise the recommender can predict which items the user will select, how the user will rate these items, or how the user will rank these items considering his/her personal preferences.

1.2.1.1 Item prediction

Predicting which items the user will select is sometimes called the item prediction problem. For this item prediction problem, each recommendation can be classified into one of the four following categories [29]:

- True Positive (TP, an interesting item that is recommended to the user)
- True Negative (TN, an uninteresting item that is not recommended to the user)
- False Negative (FN, an interesting item that is not recommended to the user)
- False Positive (FP, an uninteresting item that is recommended to the user)

Based on this classification, various accuracy metrics for the item prediction problem have been proposed such as precision, recall, and the F-score (or also called the F1-measure) [2]. Precision measures the fraction of recommendations that were successful, i.e., the fraction of recommended items that were selected by the user (in the test set) or in other words, the ratio of the number of TP recommendations and the sum of the TP an FP recommendations.

$$Precision = \frac{TP}{TP + FP} \tag{1.1}$$

Recall or the True Positive Rate (TPR) stands for the fraction of selected items that were recommended by the system or in other words, the ratio of the number of TP recommendations and the sum of the TP an FN recommendations.

$$Recall = TPR = \frac{TP}{TP + FN} \tag{1.2}$$

Typically, a trade-off has to be made between precision and recall: while allowing longer recommendation lists typically improves recall, it is also likely to reduce the precision. Therefore, the accuracy is often expressed by means of the F-score, i.e., the harmonic mean of precision and recall, which is the ratio of two times the number of TP recommendations and the sum of the FN, FP, and two times the TP recommendations.

$$F - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \tag{1.3}$$

The curve comparing the True Positive Rate, TPR (on the vertical axis), to the False Positive rate, FPR (on the horizontal axis), is known as the Receiver Operating Characteristic[7] or ROC curve, which is commonly used to evaluate the accuracy in a visual way.

$$FPR = \frac{FP}{FP + TN} \tag{1.4}$$

[7] This is a reference to their origins in signal detection theory.

1.2.1.2 Rating prediction

Predicting the ratings that users give to an item is often called the rating prediction problem. Given an active user u and his/her past rating behaviour, the recommender should be able to generate a prediction, p_{uc}, of this user's ratings for any unrated content item, c. The rating prediction accuracy is evaluated by comparing this prediction of the user's rating with the corresponding actual rating for the content item given by the user, r_{uc}. These actual ratings are only known in the evaluation phase because they are hidden for the recommender during the calculation of the recommendations in the offline experiment, or because they were obtained through a user study or online evaluation.

The Root Mean Square Error (RMSE) is perhaps the most popular metric for the rating prediction problem, partly because of the Netflix Prizein which it was used to evaluate the recommendation algorithms. Given the test set Γ of user-content item pairs, (u, c) for which the true ratings are known, the RMSE between the predicted ratings, p_{uc}, and the actual ratings, r_{uc}, is calculated as [2]:

$$RMSE = \sqrt{\frac{1}{|\Gamma|} \sum_{(u,c)\in\Gamma} (p_{uc} - r_{uc})^2} \tag{1.5}$$

Here, $|\Gamma|$ stands for the cardinality of the test set. A popular alternative metric is the Mean Absolute Error (MAE), which is calculated as:

$$MAE = \frac{1}{|\Gamma|} \sum_{(u,c)\in\Gamma} |p_{uc} - r_{uc}| \tag{1.6}$$

Compared to the MAE, the RMSE disproportionately penalizes larger prediction errors. These accuracy metrics are intended for non-binary rating mechanisms. In case of a binary rating mechanism with a binary prediction (e.g., a like/dislike system), the MAE is reduced to the misclassification rate, i.e., the ratio of the number of incorrect predictions and the total number of predictions. Then, the RMSE is the square root of this misclassification rate.

1.2.1.3 Ranking prediction

Ranking metrics evaluate the recommender in terms of ranking the content items according to the user's personal preferences. For this category of metrics, the particular set of items selected by the user and the exact rating value of the items are less important but rather the ordering of the items according to the user's preferences. To evaluate the recommendations with respect to a reference ranking (i.e., the ranking of the content items according to the user), it is first necessary to obtain such a reference ranking. In cases where only item selection or item usage is monitored, it is difficult to obtain a reference ranking. In cases were explicit, personal

ratings for items are available, the rated items can be ranked in decreasing order of the ratings, with ties for the items that received the same rating [2].

The normalized Discounted Cumulative Gain (nDCG) is a standard Information Retrieval (IR) measure [30] that can be used to evaluate the ranking of items in a recommendation list [31]. So, each user's recommendation list is a ranked list of n content items, $c_1, c_2, \ldots, c_n$, ordered according to their prediction score[8], i.e., the score of the algorithm that estimates the interest of the user in the item. For each user u, the accuracy of his/her recommendation list can be assessed based on his/her true ratings r in the test set using the Discounted Cumulative Gain (DCG) at rank n, which is computed as:

$$DCG_n^u = r_{uc_1} + \sum_{i=2}^{n} \frac{r_{uc_i}}{log_2(i)} \tag{1.7}$$

Here, r_{uc_i} stands for the true rating of user u for content item c, ranked in position i of the recommendation list.

The normalized DCG, nDCG, is calculated by the ratio of the DCG and the maximum DCG:

$$nDCG_n^u = \frac{DCG_n^u}{maxDCG_n^u} \tag{1.8}$$

where maxDCG stands for the maximum value that the DCG can get by the optimal selection and ordering of the n content items in the recommendation list $c_1, c_2, \ldots, c_n$. The optimal selection and ordering of the content items corresponds to the selection of the items with the highest true ratings of the user, ordered according to these true ratings.

The calculation of the nDCG relies on the assumption that the true ratings of the user are available for the recommended items. However in most cases, the test set contains only part of the items of the recommendation list. As solution to this, Baltrunas et al. suggested to compute the nDCG on all the rated items in the test set of the user, sorted according to the ranking computed by the recommendation algorithm [31]. Using this approach, the nDCG is calculated on the projection of the recommendation list on the test set of the user. E.g., suppose that $rec = [A, H, I, K, B, M, P, X]$ is the ordered lists of recommended items for user u, and that his/her test set contains ratings for the following seven items $test = \{Z, X, B, L, I, M, A\}$. In this case, the nDCG is computed on the ordered list $rec_{projection} = [A, I, B, M, X]$. After calculating the nDCG for each individual user, the average (arithmetic mean) nDCG over all users of the data set can be calculated as an overall measure of efficiency. This average nDCG ranges between 0 and 1; and higher values indicate a more accurate ranking of the items in the recommendation list. In addition to nDGC, different alternative ranking metrics have

[8]The prediction score is sometimes also called the recommendation score.

been proposed in literature such as the Normalized Distance-based Performance Measure (NDPM) [32] and the R-Score [33].

1.2.2 Diversity

Frequently, the recommendation lists that are presented to the users contain a lot of similar items. On Amazon.com, for example, on the webpage of a book by Robert Heinlein, users receive a recommendation list full of all of his other books [34]. Indeed, recommendation algorithms can trap users in a "similarity hole", only giving exceptionally similar suggestions [34].

Accuracy metrics cannot see this problem because they are designed to judge the accuracy of the individual recommended items; they do not judge the content of entire recommendation lists. Therefore, an additional quality metric measuring the diversity in the recommendation list is required.

The most explored method for measuring diversity in the recommendation list uses *item-item similarity*. This item-item similarity is typically calculated based on the item content [2]. Then, the diversity of the list can be measured by calculating the sum, average, minimum, or maximum distance between item pairs. Alternatively, we could measure the diversity for each item that is added to the recommendation list as the new item's diversity from the items already in the list [2, 35].

Diversity in a recommendation list is important, also in the context of group recommendations as discussed in Chapter 4. However, it is an additional quality metric, next to accuracy, and it cannot be evaluated as a stand-alone measure, since recommendations that are more diverse, might be less accurate.

1.2.3 Coverage

The coverage of a recommender system is a measure of the domain of items over which the system can make recommendations [36]. In literature, the term coverage is mainly associated with two concepts: (1) the percentage of items for which the system is able to generate a recommendation, i.e., *prediction coverage*, and (2) the percentage of the available items which effectively are ever recommended to a user, i.e., *catalog coverage* [36, 37]. In this research (especially Chapter 4), we focus on this second interpretation of coverage, thereby providing an answer to the question: "What percentage of the available items does the recommender system recommend to users?". As a result, coverage is a metric that is especially important for the system owner and less interesting for the users. Preferably as much content items as possible are reachable through the recommendations (i.e., show up in someone's recommendation list), thereby suggesting not only the same popular items to all users, but also more niche items from the long tail matching users' specific preferences. Coverage must be measured in combination with accuracy,

so good recommenders should not be tempted to raise coverage by making bogus predictions for every item in the system catalog [36].

1.2.4 Serendipity

Recommender systems might produce recommendations that are highly accurate and have reasonable diversity and coverage - and yet that are useless for practical purposes [36]. E.g., a shopping cart recommender for a grocery store might suggest potatoes to any shopper who has not yet selected them. Statistically, almost everyone buys potatoes at the grocery store; so this recommendation is highly accurate in predicting the user's purchases. However, almost everyone who shops at the grocery store has bought potatoes in the past, and knows whether or not (s)he wants to purchase more potatoes. So, the shopper has already made a concrete decision whether or not to purchase potatoes, and will therefore not be influenced by the recommendation for potatoes. These obvious recommendations are well known to the users and do not give any new information. Much more valuable are recommendations for new products or products the customer has never heard of, but would love.

Therefore, serendipity is a very desirable quality attribute of a recommendation. A serendipitous recommendation helps the user find a surprisingly interesting item (s)he might not have otherwise discovered [36]. Serendipity is a measure of how surprising the successful recommendations are [2]. Like diversity and coverage, serendipity has to be balanced with accuracy, since some recommendations, such as random suggestions, might be very surprising but not relevant for the user. So, serendipity is a measure of the amount of relevant information that is new to the user in a recommendation.

Although accuracy metrics are well known and generally accepted in the domain of recommender systems, a metric for evaluating the serendipity of a recommendation list is still an open problem. Since serendipity is a measure of the degree to which the recommendations are presenting items that are both surprising and attractive to the users, designing a metric to measure serendipity is difficult [36]. So although serendipity can best be assessed by user studies, various researchers have proposed metrics to estimate the serendipity in an offline evaluation [2, 38].

Murakami et al. [38] proposed a metric for measuring the serendipity of a recommendation list by means of the concept unexpectedness. Their metric is based on the idea that the unexpectedness is low for easy-to-predict items originating from a primitive recommender, and high for difficult-to-predict items coming from a more advanced recommender. Accordingly, the unexpectedness of a suggested item is estimated based on the difference between the confidence of the advanced recommender in the suggested item and the confidence of the primitive recommender in that suggested item. Unfortunately, the results obtained by this metric depend on the implementation of the primitive recommender and the resemblance

between the primitive and the advanced recommender. Therefore, Murakami et al. introduced three possible alternatives for the primitive recommender, providing three different values for the serendipity. Because of these drawbacks, we did not use the serendipity metric of Murakami et al. in the experiments of Chapter 4. In contrast, Shani and Gunawardana proposed a metric for the serendipity without a dependency on a primitive recommender [2], which was therefore adopted in Chapter 4 to assess the serendipity of group recommendations.

1.2.5 Novelty

Novelty is a metric that is closely related to serendipity. Serendipity measures how surprising an interesting recommendation is, given the user's past consumption behaviour, and is often assessed as the degree to which the recommended items differ from the content items that the user consumed in the past [2]. The novelty of a recommendation refers to the information value of the suggestion. A novel recommendation is a suggestion for an item that the user was unaware of, and that is discovered by exploring the recommendations. E.g., if a user did not know about the release of a new movie with his/her favorite actor, then this item is a novel recommendation. However the serendipity of this item might be rather low: this recommendation is not unexpected since the user has probably already watched multiple movies of the same genre with this actor.

Unfortunately, users do not report all items they are aware of (e.g., through a rating); they tend not to report items that they were indifferent for (i.e., items that they would give three stars on a 5-point scale star-rating mechanism) [2]. As a result, the novelty of a recommendation is difficult to measure in an offline evaluation. We can assume that popular items are generally known, but a thorough analysis of the novelty has to be performed through a user study. Through a user study, test subjects can be asked whether or not they are familiar with the recommended items.

1.2.6 Trust

The user's trust in the recommender system is an important metric for a successful service. Trust refers to the extent to which users believe they will truly like the recommendations. Regrettably, trust cannot be evaluated through an offline evaluation. In an online evaluation, we can assume that the number of recommendations that were followed is associated with the user's trust in the system. And in user studies, the test subjects can be asked whether they trust the recommendations.

1.2.7 Utility

Most e-commerce websites evaluate a recommender system in terms of the (increased) sales or the profit they make. So measuring the utility of the recommendations to the website owner is straightforward. In contrast, measuring the utility of the recommendation to the users is more difficult. Through online evaluations, the utility can be measured for the service with and without recommendations. In user studies, the utility can be assessed by a questionnaire or an interview.

1.2.8 User satisfaction

The user satisfaction refers to the overall experience that users have with a recommender system. This metric covers to some extend all previously discussed qualitative aspects, and can thus be considered as the final goal in the optimization of a recommender system. Measuring the user satisfaction is not possible in an offline evaluation; it has to be assessed via a questionnaire or an interview. In Chapter 3, the relation between the user satisfaction and other qualitative aspects of the recommender system is investigated by means of a user study.

1.2.9 Confidence

The confidence of a recommender system refers to the system's trust in its own predictions. The recommendation algorithm can report this confidence in a specific prediction next to the recommendation. This confidence metric can be used to remove recommendations with a low confidence value, thereby introducing a trade-off between confidence and coverage.

1.2.10 Risk

In some cases, the recommendations may be associated with a potential risk [2], e.g., an e-commerce website that allows purchases to be returned at the expense of the website. In this scenario, recommending bad items incurs a cost of delivery (back and forth).

1.2.11 Privacy

Some users may feel suspicious about recommender systems because of their privacy. Indeed, recommender system create a personal profile, track the user's behaviour, and infer personal preferences. These are sensitive data about the user, that have to be stored and processed securely to protect the privacy of the users. Also recommendations generated by CF algorithms should not release any personal information of other individuals in the system.

1.2.12 Robustness

Robustness is a characteristic of the recommender system that indicates how stable the recommendations are in case that fake information is passed. Stakeholders might create fake user profiles and provide fake ratings in order to increase personal profit. E.g., an hotel owner can pretend to be a customer and provide a high rating to his/her own hotel, or a low rating to a competing hotel, in order to obtain a higher ranking in the recommendation list of the travel agency website.

1.2.13 Adaptivity

For most (online) content delivery services, trends in user behaviour may be noticeable. The popularity and attractiveness of content items may shift over time: new items become interesting and old items fall into oblivion. For some use cases such as news or events, previously popular items may even become useless after a certain period of time. Therefore, it is important to take changes in information and user behaviour into account to retain a certain level of accuracy. Adaptivity is the characteristic of recommender systems indicating how fast the recommendations adapt to changes in items or trends.

1.2.14 Scalability

In many cases, recommender systems have to process huge amounts of data and thereby require a considerable amount of resources. So during implementation, trade-offs have often to be made between different resources such as computation power and memory. To keep the required resources within reasonable limits, scalability is a desired quality aspect of recommender systems. Scalability can be measured by monitoring the required resources for an increasing amount of user, items, or consumptions. In many practical scenarios, some accuracy has to be substituted for scalability, thereby indicating the trade-off between scalability and accuracy.

1.3 Group recommendations

1.3.1 Motivation and context

Recommender systems can help users to find the most interesting products or content thereby addressing the information overload problem of (online) services. Personal preferences are extracted from the users' historical feedback in order to suggest each user the most suitable items. Although the majority of the currently deployed recommender systems are designed to generate personal suggestions for individual users, in many cases content is selected and consumed by groups of

users rather than by individuals. E.g., movies or TV shows are often watched in a family context, people go to restaurants, bars, and (cultural) events with their friends, and choosing a holiday destination is mostly a joint decision of the travel group. These scenarios introduce the need for discovering the most appropriate group recommendation strategies for video-on-demand services, event websites, services providing information about points-of-interest, travel agencies, etc.

The first scientific publications regarding recommender systems for groups date from the late nineties [39]. From then, many researchers have already investigated how the current state-of-the-art recommendation algorithms can be adapted in order to generate group recommendations. In literature, group recommendations have mostly been generated either by aggregating the users' individual recommendations into recommendations for the whole group (aggregating recommendations) or by aggregating the users' individual preference models into a preference model of the group (aggregating preferences) [40]. In this dissertation, we refer to these strategies as *aggregation strategies*.

The first aggregation strategy (aggregating recommendations) generates recommendations for each individual user using a general recommendation algorithm. Subsequently, the recommendation lists of all group members are aggregated into a group recommendation list which (hopefully) satisfies all group members. Different approaches to aggregate the recommendation lists have been proposed during the last decade. Most of them make a decision based on the algorithm's prediction score, i.e., a prediction of the user's rating score for the recommended item. The higher the prediction score is, the better the match between the user's preferences and the recommended item. Aggregating the users' individual recommendations into group recommendations has some advantages. For instance, the resulting recommendations can be directly linked to the individual recommendations (i.e., recommendations for a single user), which makes them easy to explain based on the explanations of the traditional recommender [41]. Conversely, the link between the group recommendations and the individual recommendations makes it less likely to identify unexpected, surprising items [7].

The second aggregation strategy (aggregating preferences) combines the users' preferences into group preferences. This way, the opinions and preferences of individual group members constitute a group preference model reflecting the interests of all members. In literature, different approaches have been proposed to aggregate the members' preferences, but still no consensus exists about the optimal solution [31, 42]. After aggregating the members' preferences, the group's preference model is treated as a pseudo user in order to produce recommendations for the group using a traditional recommendation algorithm. Compared to aggregating the individual recommendation lists, aggregating the users' preferences increases the chance of finding serendipitously valuable recommendations. On the other hand, aggregating the preferences may lead to group suggestions that lie outside

the range of any individual recommendation list, which may be disorienting to the users and difficult to explain [41].

In this dissertation, we refer to the methods that aggregate the individual recommendation lists into group recommendations or combine the group members' preferences into a group preference model as *(data) aggregation methods*.

1.3.2 Existing systems and related work on group recommendations

From the late nineties, many group recommender systems have been proposed in literature. In this section, we provide an overview of the existing group recommenders for various domains of items such as music, TV-shows and movies, touristic points-of-interest, web pages, etc.

In 1998 MusicFX was presented, a system to select background music for a group of people working out in a fitness centre [39]. Based on the preferences of the people, the system constructs a group profile (by aggregating the preferences) and selects a music channel including some randomness in the choice procedure to ensure variety. According to a quantitative assessment, the vast majority of fitness centre members who were involved in this trial were pleased with the group recommendations. Another music recommender for groups of users in the same environment is Flytrap [43]. Based on the music people listen to on their computers, Flytrap automatically constructs a soundtrack that tries to please everyone in the room. The system detects the presence of people in the room by the radio frequency ID badge of every user and generates recommendations by aggregating the votes of all users (cfr. aggregating preferences strategy). Adaptive Radio is another system that selects music to play in a shared environment [44]. This recommender discovers what a user does not like instead of what the user does like. Based on these (aggregated) negative preferences, music suggestions are produced that are acceptable for all members of a group.

In the domain of movies, PolyLens is an extension of MovieLens that enables recommendations for groups [7]. This recommender system uses CF to recommend movies for users based on the users' star ratings. Polylens uses an algorithm that merges the users' recommendation lists (cfr. aggregating recommendations strategy), thereby avoiding movies that any member of the group has already rated (and therefore seen). Polylens allows users to create and manage their own groups in order to receive group recommendations next to the traditional individual recommendations. Both survey results and observations of user behaviour proved that group recommendations are valuable and desirable for the users. They also revealed that users are willing to share their personal recommendations with the group, thereby trading some privacy for group recommendations. In the context of recommendations for TV-content, the Family Interactive TV system filters TV

programs and creates an adaptive programming guide according to the different viewers' preferences [45]. The group recommendations of this system are based on implicit relevance feedback that is assessed through the actual program the viewer has chosen for watching. Also in the context of watching TV in group, three alternative strategies for generating group recommendations are analysed and compared: a common group profile, aggregating recommendations, and aggregating preferences [46]. A common group profile can be considered as a virtual user of the system, representing all group members. Through a common group profile, users cannot evaluate content individually, since they have to give ratings or provide feedback for the group as a whole. The aggregating preferences strategy is chosen as optimal solution for their TV recommender. Their data aggregation method is based on total distance minimization, which guarantees that the merged result is close to most users' preferences. The evaluation results proved that the recommendation strategy is effective for multiple viewers watching TV together and appropriately reflects the preferences of the majority of the members within the group. Beside video watching in the home environment, multimedia content is often viewed by users on the move. Therefore, an adaptive vehicular multimedia system has been developed to personalize the multimedia based on the aggregation of the preferences of groups of passengers travelling together in buses, trains, and airplanes [47] (cfr. aggregating preferences strategy).

Many group recommender systems for points-of-interest (POI's) such as touristic attractions, restaurants, hotels, etc. have been proposed in literature. The Pocket Restaurant Finder provides restaurant recommendations for groups that are planning to go out eating together. The application can use the physical location of the kiosk or mobile device on which it is running, thereby taking into account the position of the people on top of their culinary preferences. Users have to specify their preferences regarding the cuisine type, restaurant amenities, price category, and ranges of travel time from their current location on a 5-point rating scale. When a group of people is gathered together, the Pocket Restaurant Finder pools these preferences together (cfr. aggregating preferences strategy) and presents a list of potential restaurants, sorted in order of expected desirability for the group using a content-based algorithm [48]. Intrigue is a group recommender system for touristic places which considers the characteristics of subgroups such as children or disabled and addresses the possibly conflicting preferences within the group. In this system, the preferences of these heterogeneous subgroups of people are managed and combined by using a group model in order to identify solutions satisfactory for the group as a whole [49]. Also in the context of touristic activities, the Travel Decision Forum is an interactive system that assists in the decision process of a group of users planning to take a vacation together [50]. The mediator of this system directs the interactions between the users thereby helping the members of the group to agree on a single set of criteria that are to be applied in the making

of a decision. This recommender takes into account people's preferences regarding various characteristics such as the facilities that are available in the hotel room, the sightseeing attractions in the surrounding area, etc [51]. An alternative recommender system for planning a vacation is CATS (Collaborative Advisory Travel System) [52]. It allows a group of users to simultaneously collaborate on choosing a skiing holiday package which satisfies the group as a whole. This system has been developed around the DiamondTouch interactive tabletop, which makes it possible to develop a group recommender that can be physically shared between up to four users. Recommendations are based on the group profile, which is a combination of individual personal preferences (cfr. aggregating preferences strategy). The last example in the domain of POI's is Group Modeller, a group recommender that provides information about museums and exhibits for small groups of people [53]. This recommender system creates group models from a set of individual user models.

Although Web browsing is usually a solitary activity, like most of today's desktop applications, various research initiatives have tried to assist a group of people in browsing by suggesting new material likely to be of common interest. Let's Browse is an extension of a single user browser that recommends web pages to a group of people using a content-based algorithm [54]. This recommender system estimates the interests of the users by analysing the words of the visited web pages of each individual and of the groups. The system uses a simple linear combination of the profiles of each user (cfr. aggregating preferences strategy), so that the recommendation is the page that scored the best in the combined profile. Other interesting features of Let's Browse are the automatic detection of the presence of users, the dynamic display of the user profiles, and the explanation of recommendations. I-SPY is a collaborative, community-based search engine that recognizes the implicit preferences of communities of searchers and personalizes the search results [55]. This personalized search engine offers potential improvements in search performance, especially in certain situations where communities of searchers share similar information needs and use similar queries to express these needs.

Another use case of group recommendations is a recipe recommender for families [40]. Since all family members typically eat a joint meal at least once a day, choosing a recipe and consuming the food are good examples of a group activity. In the context of this recipe recommender, the aggregating preferences strategy and the aggregating recommendations strategy were compared. An evaluation with a number of families showed that for users with a low density profile (i.e., containing a small number of consumptions), the aggregated recommendation lists yield slightly better results than the aggregated preferences. For users with a higher density profile on the other hand, the recommendations obtained by aggregating the users' profiles showed to be more accurate, than the aggregated recommen-

dation lists. This recommender system is based on CF and the individual data of group members is aggregated in a weighted, domain-dependent manner, such that the weights reflect the observed interaction of the group members. As was already remarked by other researchers, this is only one type of recommendation algorithm and one of the many possible approaches for aggregating preferences or recommendation lists [31]. So, an extensive comparison of the two aggregation strategies is still missing in literature.

Research regarding the strategy that aggregates the individual recommendation lists into a list of group recommendations (cfr. aggregating recommendations strategy) has demonstrated that the influence of the data aggregation method is limited (i.e., the way in which recommendation lists for individual users are aggregated into a group recommendation list) [31]. A comparison of the group recommendation lists generated using four commonly used aggregation methods showed similar results in terms of accuracy for all methods. This study has also compared the accuracy of these group recommendations with individual recommendations (i.e., recommendations for a single user). For small groups, the group recommendations showed to be only slightly less effective than the individual recommendations, whereas for larger groups, the group recommendations are significantly inferior than the individual recommendations. If the groups are selected in such a way that the members have preferences that are quite similar, the study showed that the effectiveness of group recommendations does not necessarily decrease when the group size increases.

In chapter 4, we thoroughly investigate the two different strategies to generate group recommendations by comparing the accuracy of the group recommendations for various sizes of the group. Besides, the influence of the similarity between group members on the accuracy of the group recommendations is investigated. In contrast to existing research [31, 40], our work goes further by comparing group recommendations generated by using various traditional recommendation algorithms. The results show that the best strategy for generating group recommendations is depending on the recommendation algorithm that is used to generate suggestions for individuals. For all algorithms, the accuracy evaluation indicates that the more alike the users of a group are, the more effective the group recommendations are. However being accurate is not enough for a recommendation list [34]; also other characteristics like diversity, coverage, and serendipity are essential for a valuable list of suggestions. Therefore, our research also considers these additional quality metrics, whereas other studies merely focus on accuracy as the only metric for evaluating (group) recommendations [31, 40].

1.4 Conclusion

Recommender systems are an active research area with a growing importance and have become a common tool for online services in recent years. Many different techniques and algorithms have been proposed to generate personalized recommendations matching the preferences of the users. Nevertheless, evaluating recommender systems remains tricky since no standardized evaluation procedure exists. Moreover, recommendations, and especially group recommendations, are generally only evaluated in terms of accuracy, thereby neglecting other qualitative characteristics such as diversity and serendipity. This dissertation contributes to the research area of recommender systems by evaluating different state-of-the-art recommendation algorithms from a user perspective. In addition, this dissertation presents a thorough evaluation of group recommendations in terms of different qualitative attributes and an innovative strategy for generating group recommendations.

References

[1] P. Resnick and H. R. Varian. *Recommender systems*. Communications of the ACM, 40(3):56–58, March 1997.

[2] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor. *Recommender Systems Handbook*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

[3] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.

[4] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. *Using collaborative filtering to weave an information tapestry*. Communications of the ACM, 35(12):61–70, 1992.

[5] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. *GroupLens: an open architecture for collaborative filtering of netnews*. In Proceedings of the 1994 ACM conference on Computer Supported Cooperative Work, CSCW '94, pages 175–186, New York, NY, USA, 1994. ACM.

[6] J. A. Konstan, J. Riedl, A. Borchers, and J. L. Herlocker. *Recommender systems: A grouplens perspective*. In Proceedings of the AAAI Workshop on Recommender Systems 1998, AAAI Technical Report WS-98-08, pages 60–64, 1998.

[7] M. O'Connor, D. Cosley, J. A. Konstan, and J. Riedl. *PolyLens: a recommender system for groups of users*. In Proceedings of the seventh European Conference on Computer Supported Cooperative Work, ECSCW'01, pages 199–218, Norwell, MA, USA, 2001. Kluwer Academic Publishers.

[8] R. Burke. *Hybrid Recommender Systems: Survey and Experiments*. User Modeling and User-Adapted Interaction, 12(4):331–370, 2002.

[9] T. Mahmood and F. Ricci. *Towards Learning User-Adaptive State Models in a Conversational Recommender System*. In 15th Workshop on Adaptivity and User Modeling in Interactive Systems, 2007.

[10] D. Bridge, M. H. Göker, L. McGinty, and B. Smyth. *Case-based recommender systems*. The Knowledge Engineering Review, 20(03):315–320, 2005.

[11] F. Ricci, D. Cavada, N. Mirzadeh, A. Venturini, D. Fesenmaier, K. Wöber, H. Werthner, et al. *Case-based travel recommendations.* Destination recommendation systems: behavioural foundations and applications, pages 67–93, 2006.

[12] A. Aamodt and E. Plaza. *Case-based reasoning: foundational issues, methodological variations, and system approaches*. AI Communications, 7(1):39–59, 1994.

[13] A. Felfernig and R. Burke. *Constraint-based recommender systems: technologies and research issues*. In Proceedings of the 10th international conference on Electronic commerce, ICEC '08, pages 3:1–10, New York, NY, USA, 2008. ACM.

[14] R. Sinha and K. Swearingen. *Comparing Recommendations Made by Online Systems and Friends*. In Proceedings of the DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries, 2001.

[15] J. Golbeck. *Generating predictive movie recommendations from trust in social networks*. In Proceedings of the 4th international conference on Trust Management, iTrust'06, pages 93–104. Springer-Verlag Berlin Heidelberg, 2006.

[16] M. Pazzani and D. Billsus. *Learning and Revising User Profiles: The Identification of Interesting Web Sites*. Machine Learning, 27(3):313–331, 1997.

[17] T. Joachims. *Text categorization with Support Vector Machines: Learning with many relevant features*. In C. Ndellec and C. Rouveirol, editors, Machine Learning: ECML-98, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer Berlin Heidelberg, 1998.

[18] J. Rocchio. *Relevance Feedback in Information Retrieval*. SMART Retrieval System Experimens in Automatic Document Processing, 1971.

[19] D. Billsus, M. J. Pazzani, and J. Chen. *A learning agent for wireless news access*. In Proceedings of the 5th international conference on Intelligent user interfaces, IUI '00, pages 33–36, New York, NY, USA, 2000. ACM.

[20] M. Hearst, S. Dumais, E. Osman, J. Platt, and B. Scholkopf. *Support vector machines*. IEEE Intelligent Systems and their Applications, 13(4):18–28, 1998.

[21] P. Pu, L. Chen, and R. Hu. *A user-centric evaluation framework for recommender systems*. In Proceedings of the fifth ACM conference on Recommender systems, RecSys '11, pages 157–164, New York, NY, USA, 2011. ACM.

[22] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. *GroupLens: applying collaborative filtering to Usenet news*. Communications of the ACM, 40(3):77–87, March 1997.

[23] G. Linden, B. Smith, and J. York. *Amazon.com recommendations: item-to-item collaborative filtering*. Internet Computing, IEEE, 7(1):76–80, 2003.

[24] M. Grcar, B. Fortuna, D. Mladenic, and M. Grobelnik. *kNN Versus SVM in the Collaborative Filtering Framework*. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna, editors, Data Science and Classification, Studies in Classification, Data Analysis, and Knowledge Organization, pages 251–260. Springer Berlin Heidelberg, 2006.

[25] R. Bell, Y. Koren, and C. Volinsky. *Modeling relationships at multiple scales to improve accuracy of large recommender systems*. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '07, pages 95–104, New York, NY, USA, 2007. ACM.

[26] Z. Huang, H. Chen, and D. Zeng. *Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering*. ACM Transactions on Information Systems, 22(1):116–142, January 2004.

[27] K. De Moor, T. De Pessemier, P. Mechant, C. Courtois, L. De Marez, A. Juan, and L. Martens. *Users' (Dis)satisfaction with the personalTV application: Combining objective and subjective data*. Computers in Entertainment, 9(3):18:1–18:22, November 2011.

[28] E. Campochiaro, R. Casatta, P. Cremonesi, and R. Turrin. *Do Metrics Make Recommender Algorithms?* In Proceedings of the 2009 International Conference on Advanced Information Networking and Applications Workshops,

WAINA '09, pages 648–653, Washington, DC, USA, May 2009. IEEE Computer Society.

[29] P. Cremonesi, R. Turrin, E. Lentini, and M. Matteucci. *An Evaluation Methodology for Collaborative Recommender Systems*. In Proceedings of the 2008 International Conference on Automated solutions for Cross Media Content and Multi-channel Distribution, AXMEDIS '08, pages 224 –231, November 2008.

[30] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to information retrieval*. Cambridge Univ. Press, New York, NY, USA, 2008.

[31] L. Baltrunas, T. Makcinskas, and F. Ricci. *Group recommendations with rank aggregation and collaborative filtering*. In Proceedings of the fourth ACM conference on Recommender systems, RecSys '10, pages 119–126, New York, NY, USA, 2010. ACM.

[32] Y. Y. Yao. *Measuring retrieval effectiveness based on user preference of documents*. Journal of the American Society for Information Science, 46(2):133–145, 1995.

[33] J. S. Breese, D. Heckerman, and C. Kadie. *Empirical analysis of predictive algorithms for collaborative filtering*. In Proceedings of the fourteenth conference on Uncertainty in artificial intelligence, UAI'98, pages 43–52, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[34] S. M. McNee, J. Riedl, and J. A. Konstan. *Being accurate is not enough: how accuracy metrics have hurt recommender systems*. In CHI '06 extended abstracts on Human factors in computing systems, CHI EA '06, pages 1097–1101, New York, NY, USA, 2006. ACM.

[35] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. *Improving recommendation lists through topic diversification*. In Proceedings of the 14th international conference on World Wide Web, WWW '05, pages 22–32, New York, NY, USA, 2005. ACM.

[36] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. *Evaluating collaborative filtering recommender systems*. ACM Transactions on Information Systems, 22(1):5–53, 2004.

[37] M. Ge, C. Delgado-Battenfeld, and D. Jannach. *Beyond accuracy: evaluating recommender systems by coverage and serendipity*. In Proceedings of the fourth ACM conference on Recommender systems, RecSys '10, pages 257–260, New York, NY, USA, 2010. ACM.

[38] T. Murakami, K. Mori, and R. Orihara. *Metrics for Evaluating the Serendipity of Recommendation Lists*. In K. Satoh, A. Inokuchi, K. Nagao, and T. Kawamura, editors, New Frontiers in Artificial Intelligence, volume 4914 of *Lecture Notes in Computer Science*, pages 40–46. Springer Berlin Heidelberg, 2008.

[39] J. F. McCarthy and T. D. Anagnost. *MusicFX: an arbiter of group preferences for computer supported collaborative workouts*. In Proceedings of the 1998 ACM conference on Computer Supported Cooperative Work, CSCW '98, pages 363–372, New York, NY, USA, 1998. ACM.

[40] S. Berkovsky and J. Freyne. *Group-based recipe recommendations: analysis of data aggregation strategies*. In Proceedings of the fourth ACM conference on Recommender systems, RecSys '10, pages 111–118, New York, NY, USA, 2010. ACM.

[41] J. L. Herlocker, J. A. Konstan, and J. Riedl. *Explaining collaborative filtering recommendations*. In Proceedings of the 2000 ACM conference on Computer Supported Cooperative Work, CSCW '00, pages 241–250, New York, NY, USA, 2000. ACM.

[42] J. Masthoff. *Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers*. User Modeling and User-Adapted Interaction, 14:37–85, 2004.

[43] A. Crossen, J. Budzik, and K. J. Hammond. *Flytrap: intelligent group music recommendation*. In Proceedings of the 7th international conference on Intelligent user interfaces, IUI '02, pages 184–185, New York, NY, USA, 2002. ACM.

[44] D. L. Chao, J. Balthrop, and S. Forrest. *Adaptive radio: achieving consensus using negative preferences*. In Proceedings of the 2005 international ACM SIGGROUP conference on supporting group work, GROUP '05, pages 120–123, New York, NY, USA, 2005. ACM.

[45] D. Goren-Bar and O. Glinansky. *Family Stereotyping - A Model to Filter TV Programs for Multiple Viewers*. In Proceedings of the 2nd Workshop on Personalization in Future TV, May 2002.

[46] Z. Yu, X. Zhou, Y. Hao, and J. Gu. *TV Program Recommendation for Multiple Viewers Based on user Profile Merging*. User Modeling and User-Adapted Interaction, 16:63–82, March 2006.

[47] Y. Zhiwen, Z. Xingshe, and Z. Daqing. *An adaptive in-vehicle multimedia recommender for group users*. In IEEE 61st Vehicular Technology Conference, 2005. VTC 2005-Spring, volume 5, pages 2800–2804, May-June 2005.

[48] J. McCarthy. *Pocket RestaurantFinder: A Situated Recommender System for Groups*. In Proceedings of the Workshop on Mobile AdHoc Communication at the 2002 ACM Conference on Human Factors in Computer Systems. ACM, 2002.

[49] L. Ardissono, A. Goy, G. Petrone, M. Segnan, and P. Torasso. *Tailoring the Recommendation of Tourist Information to Heterogeneous User Groups*. In S. Reich, M. Tzagarakis, and P. De Bra, editors, Hypermedia: Openness, Structural Awareness, and Adaptivity, volume 2266 of *Lecture Notes in Computer Science*, pages 228–231. Springer Berlin Heidelberg, 2002.

[50] A. Jameson, S. Baldes, and T. Kleinbauer. *Two methods for enhancing mutual awareness in a group recommender system*. In Proceedings of the working conference on Advanced visual interfaces, AVI '04, pages 447–449, New York, NY, USA, 2004. ACM.

[51] A. Jameson. *More than the sum of its members: challenges for group recommender systems*. In Proceedings of the working conference on Advanced visual interfaces, AVI '04, pages 48–54, New York, NY, USA, 2004. ACM.

[52] K. McCarthy, M. Salamo, L. Coyle, L. McGinty, B. Smyth, and P. Nixon. *CATS: A Synchronous Approach to Collaborative Group Recommendation.* In G. Sutcliffe and R. Goebel, editors, FLAIRS Conference, pages 86–91. AAAI Press, 2006.

[53] J. Kay and W. Niu. *Adapting Information Delivery to Groups of People*. In Proceedings of the Workshop on New Technologies for Personalized Information Access at the Tenth International Conference on User Modeling, 2006.

[54] H. Lieberman, N. van Dyke, and A. Vivacqua. *Let's browse: a collaborative browsing agent*. Knowledge-Based Systems, 12(8):427 – 431, 1999.

[55] B. Smyth, E. Balfe, J. Freyne, P. Briggs, M. Coyle, and O. Boydell. *Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine*. User Modeling and User-Adapted Interaction, 14:383–423, 2004.

2

Evaluating the PersonalTV service: a recommender system case study

2.1 Introduction

More and more recommender systems are being integrated with web-based platforms that suffer from information overload. By personalizing content based on user preferences, recommender systems assist in selecting relevant items on these websites. In this respect, this chapter discusses the users' satisfaction with recommendations of the 'PersonalTV' service, and chapter 3 presents a user-centric evaluation of recommendation algorithms performed on a cultural events website.

PersonalTV is an online video delivery service, consisting of a recommender system and a video player application, that has been developed for research purposes. The video player application has a desktop version, which is discussed and evaluated in this chapter, and a mobile version, which was used to study the influence of QoE on the rating behaviour in Chapter 8. The PersonalTV service enables its users to explore and watch videos from the YouTube library while it builds up a personal viewing profile in order to give personalized content suggestions.

This chapter discusses the results of the evaluation of PersonalTV by a panel of test subjects. Evaluating PersonalTV as a recommender system case study provides insights into the users' interaction behaviour and helps to understand how a recommender system is experienced and evaluated from a user point of view. The results of this study and the gathered qualitative user feedback can serve as input for improving and optimizing users' experiences with recommender systems.

2.2 Test setup

2.2.1 Goals of the study

This chapter discusses the setup and results from a panel study that aimed to evaluate a recommender system from a user point of view, thereby providing insights into the users' interaction behaviour and experience with the system[1]. Subjective self-report data were complemented with objective monitoring data and implicit feedback based on interaction behaviour. The subjective data were gathered by means of a traditional star-rating mechanism and four structured questionnaires and can be seen as explicit and subjective user feedback: users explicitly report on their experiences, preferences, etc. Additional objective data were gathered without interrupting the user's experience. In this respect, a large variety of parameters and interactions with the PersonalTV application were recorded for data-mining purposes. These interactions can be interpreted as implicit feedback and often fuel Web 2.0 applications that try to maximize the collective intelligence [1].

The goal of this research is to gain insights into the users' interaction behaviour and subjective evaluation process by triangulating the objective monitoring data, subjective explicit feedback, and implicit feedback based on interaction behaviour. Firstly, the relation between the content retrieval method, i.e., the way videos are explored (objective monitoring data) and the consumption percentage, i.e., the fraction of the video that is actually watched (implicit feedback) is explored (Section 2.3.3). Secondly, this chapter discussed the influence of the content retrieval method (objective monitoring data) on the reported satisfaction (subjective explicit feedback) (Section 2.3.4). Thirdly, the relation between the consumption percentage (implicit feedback) and the reported satisfaction (subjective explicit feedback) is investigated (Section 2.3.5). Hereby, the expected convergence of implicit and explicit feedback is tested. Finally, additional qualitative user feedback was collected in order to improve and optimize the PersonalTV service and the design of recommender systems in general.

2.2.2 Procedure

This section is dedicated to the setup and evaluation procedure of our experiment. First, the architecture and functionality of the PersonalTV service is described. Next, the recommendation algorithm is explained in detail. Afterwards, we elaborate on how test subjects are recruited for this experiment. Finally, the evaluation procedure and data collection are explained.

[1] This user study as well as the experiments regarding Quality of Experience are conducted in close cooperation with the MICT research group, Department of Communication Sciences, Ghent University.

2.2.2.1 The PersonalTV service

The PersonalTV service enables users to browse through the collection of YouTube videos: users can check the most viewed or top rated videos or can search for videos based on keywords. Furthermore, users can rate these video, thereby building a user profile with their personal evaluations and historical viewing behaviour. Based on this profile, the PersonalTV service deduces the personal video preferences of the user and generates a tailored offer of personal suggestions for unseen videos. This profile is continually updated and refined as more viewing and rating data becomes available. Consequently, the more videos a user watches and rates, the better these personal suggestions should be.

An architectural overview of the PersonalTV service is illustrated in Figure 2.1, which shows the three main components of the system: the PersonalTV video player application (client), the PersonalTV recommender system (server), and the YouTube video service.

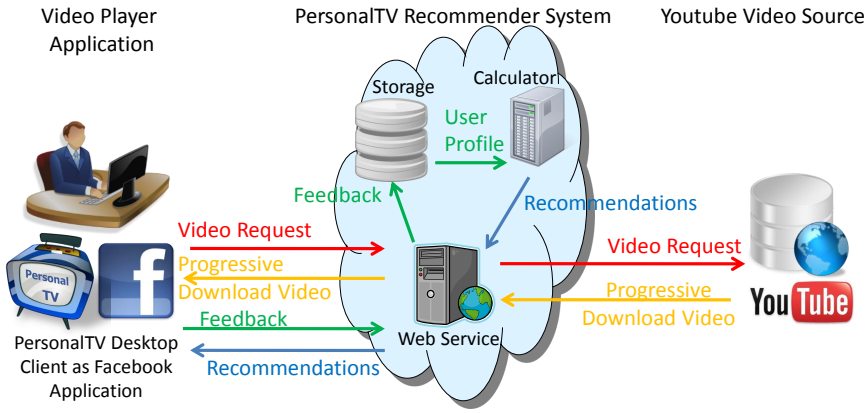


Figure 2.1: The architecture of the PersonalTV service, consisting of the video player application, the recommender system, and YouTube as video source

The PersonalTV video player application has two versions: a desktop client used to investigate the users' subjective evaluations and interaction behaviour with the video delivery system and the recommendations (as discussed in this chapter), and a mobile client used to study the influence of Quality of Experience (QoE) on the rating behaviour (as debated in Chapter 8). The desktop version of the PersonalTV video player is implemented as a Facebook application, enabling the social network functions of Facebook within PersonalTV. This way, users can easily post their favourite videos on their Facebook feed, invite their Facebook friends to try PersonalTV, and recommend videos to their Facebook friends. This integration of PersonalTV with a social network such as Facebook, was very important

to increase the visibility of PersonalTV and thereby recruiting new test subjects. Moreover, Facebook provides applications the basic profile information of users (e.g., age, gender, region) and takes care of the user authentication. The integration of PersonalTV with Facebook ensures that users can login on Facebook to authenticate themselves and do not need to create a new account for the PersonalTV service.

Figure 2.2 shows a screenshot of the PersonalTV video player application, illustrating the main features of the application. Through the PersonalTV video player application, users can explore, select, and watch YouTube videos based on various criteria. The first and the second tab offer users a view on respectively the most viewed and top rated YouTube videos within a selected period of time (today, last week, last month, or all time). The third tab gives users the opportunity to search for specific YouTube videos by providing keywords (and a period of time). The fourth tab presents the user a set of personalized suggestions for unseen videos corresponding to the user's personal preferences that are derived from the ratings. Details about the recommender system and algorithm used to calculate these suggestions are provided in Section 2.2.2.2. The fifth tab combines the functionalities of the third and fourth tab. Through this tab, users can perform a keyword-based search query, and subsequently retrieve the search results ordered by their personal preferences. This way, e.g., a query for videos with the keyword "jaguar" will yield very different results for a car enthusiast and an animal lover. By making the YouTube videos available within the PersonalTV application, users have an enormous amount of content at their disposal, in which they can certainly find videos that suits their personal preferences.

Besides browsing, selecting, and watching YouTube videos, the PersonalTV video player application gives users the opportunity to evaluate the videos through a 5-point scale star-rating mechanism, similar to the one YouTube used to have[2]. We opted for this explicit feedback method since it is adopted by many video delivery systems that provide personal recommendations such as Movielens[3], and Netflix[4]. But also in other domains, the 5-point scale star-rating mechanism is often used to gather feedback and evaluate items. E.g., TripAdvisor[5] gathers feedback about hotels and shops; and on Ebay[6] customers can evaluate vendors or vice versa, using such a star-rating mechanism. As a result, many users are familiar with this star-rating mechanism. Figure 2.3 shows a screenshot of the PersonalTV

[2] At the end of March, 2010, the YouTube website was redesigned. This resulted in a simplified website: the rating system, e.g., was changed into a 'thumbs-up, thumbs-down' scoring system.

[3] http://movielens.umn.edu/

[4] https://dvd.netflix.com/

[5] http://www.tripadvisor.com/

[6] http://www.ebay.com/

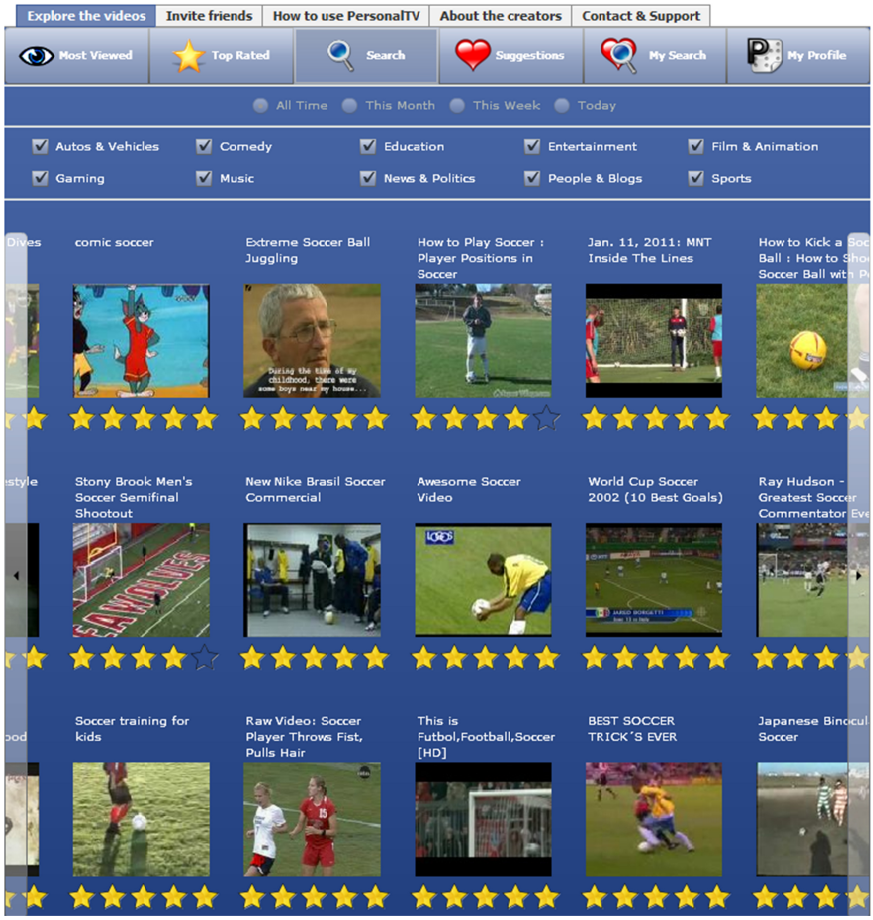


Figure 2.2: Screenshot of the PersonalTV application, showing the main features of the video player

application illustrating the star-rating mechanism as well as the functionality to recommend a video to friends and publish a video on the user's Facebook feed.

The video requests of the client application are handled by the PersonalTV web service, which sends the video via Dynamic Adaptive Streaming over HTTP (DASH), also known as progressive download, using YouTube as video source. The PersonalTV service logs these video requests, user's explicit feedback in the form of ratings, and different aspects of the user's viewing behaviour, such as, e.g., the percentage of a video that was actually watched, the number of times that the user watches a particular video, and the time of watching.

Figure 2.3: Screenshot of the star-rating mechanism of PersonalTV

These data, which are logged in the storage system of PersonalTV, are necessary to infer the user's preferences regarding video watching and create a personal user profile. Based on these user profiles, the calculator generates for all users video recommendations matching their personal preferences. The recommendations are made available via the PersonalTV web service and visible for the user through the personalized tab of the PersonalTV application (i.e., the fourth tab with the label "Suggestions"). In the same way as with the videos of the other tabs, users can browse, select, watch, and evaluate these recommended videos, thereby adjusting their personal profile in the PersonalTV recommender system. Since the preferences of the user may change over time, user behaviour and ratings are continuously logged as the PersonalTV application is used. As a result,

the personal profile of each user is continuously adjusted according to the user's feedback in order to optimally reflect the user's actual preferences.

2.2.2.2 The PersonalTV recommendation algorithm

Given the relatively small number of users of the PersonalTV service (Section 2.2.3), compared to the large number of videos that are available on YouTube, the sparse consumption matrix makes collaborative filtering algorithms less suitable for the PersonalTV service. In contrast, the availability of metadata regarding the YouTube videos (title, category, name of the uploader, and tags) enables the use of content-based algorithms. The content-based algorithm developed for PersonalTV makes optimal use of these metadata in order to generate personal suggestions.

More concretely, the algorithm makes a prediction of the user's ratings for unseen videos based on the associated metadata of the videos and the user's viewing and rating history. Subsequently, the videos with the highest prediction values are presented as personal suggestions for the user through the PersonalTV client application. The rating predictions are calculated using the personal user profile, which is composed of video metadata and previous user feedback on videos. Each user profile contains a number of monotonously increasing frequency values based on the rating behaviour of the user. More specifically, $N_u(r, f)$ denotes the frequency of occurrence that a user u, has associated a rating r, to a content item with a feature f. Such a feature is a metadata element describing the video, e.g., a tag, a category, a keyword from the title, or the name of the uploader of the video. For each video watched by the user, the metadata of the video and the star-rating are logged and used to update these frequency values in the user's personal profile.

By partitioning the frequency values of the profile $N_u(r, f)$, according to the rating r, the features in the profile f can be classified in a number of feature clouds F_r, one for every possible rating value of the system. Then, the feature cloud F_r collects all features of videos that ever received a rating r from the user, together with the frequency values $N_u(r, f)$, indicating the number of times that a video with feature f was evaluated with star-rating r by the user.

To determine whether an unseen video, or a new content item in general, is a suitable recommendation, the metadata of the new content item are compared with the user profile and the user's (future) rating for the new content item is predicted. For this comparison, the recommendation algorithm calculates the similarity between the features of the content item c, and the feature cloud F_r, which contains the features of content items that received a rating r from the user in the past. This similarity value, calculated as a sum of frequency values, indicates the resemblances between the new content item and content items that received a rating r in the past.

$$Sim_u(c, F_r) = \sum_{\forall f_i \in F_c} N_u(r, f_i) \tag{2.1}$$

Here, $Sim_u(c, F_r)$ stands for the similarity between the new content item c, and the feature cloud F_r in the profile of user u. F_c denotes the feature cloud of the new content item c, or in other words the set of features that describe this content item.

Since different features have a different frequency of occurrence in a (user-generated) content delivery system, a frequency correction factor has to be added to equation 2.1 in order to not favour the videos with only popular features. Incorporating a frequency correction factor is common practice in the context of information retrieval and text mining. But the commonly used Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme [2] was not feasible in the PersonalTV recommender, because it was not possible to obtain the frequency of a feature across all available YouTube videos. (At the time of implementation, it was not possible to query the exact number of video's with a certain feature through YouTube's API (Application Programming Interface).) Therefore, the used frequency correction factor is based on (the inverse of) the relative frequency of a feature in the user's profile containing the complete rating history of the user. As in the TF-IDF scheme, the logarithm of the frequency of the feature is used. Since uncommon or new features in the user's profile are favoured by the correction factor, they have a fair chance to have an influence on the recommendations, thereby increasing the novelty of the recommendations.

$$Sim_u(c, F_r) = \sum_{\forall f_i \in F_c} N_u(r, f_i) \cdot log \frac{N_u}{N_u(f_i)} \tag{2.2}$$

Here, N_u denotes the number of content items that user u has already rated. $N_u(f_i)$ is a subset of N_u and represents the number of content items with a feature f_i that were rated by user u. This similarity, $Sim_u(c, F_r)$, is calculated for every possible rating value in the system (i.e., r is ranging from 1 to 5 for the used 5-point scale star-rating mechanism).

Subsequently, the prediction $P_u(c)$ of the rating that user u will give to the content item c is calculated as a weighted average from these similarity values.

$$P_u(c) = \frac{1}{S} \sum_{r=r_{min}}^{r=r_{max}} Sim_u(c, F_r) \cdot r \tag{2.3}$$

In this equation, S is a normalisation factor and is defined as:

$$S = \sum_{r=r_{min}}^{r=r_{max}} Sim_u(c, F_r) \tag{2.4}$$

The concept of the algorithm is as follows. Unseen content items will probably receive a positive evaluation if these items have features that already received much positive feedback from the user (through ratings for other items with the

same features). These items will have a high similarity with high-rating feature clouds, i.e., F_4 and F_5. As a result, these videos will have a high probability of being recommended. In contrast, unseen content items will probably be negatively evaluated if these items have features that are negatively evaluated by the user in the past. These item have a high similarity with low-rating feature clouds, i.e., F_1 and F_2. Because of the user's negative experience with content items characterized by the same features, these unseen content items will not be recommended.

Content items that are listed in the user's viewing history, and therefore already have been seen by the user, are not considered as recommendations. Since the user has already explored and discovered these items, recommending the item again would not be very useful.

In Section 2.3.5, we investigate if detailed logged data of users' viewing behaviour could be used as implicit feedback to the system, and as a feasible alternative for explicit ratings. If implicit and explicit feedback actually converge, systems in which explicit ratings are not available can rely on this implicit feedback for the generation of personalized recommendations. Moreover, if explicit feedback is available, implicit feedback can be used as a complementary source of information in addition to the explicit feedback.

More concretely, the percentage of the video that is actually watched by the user is logged as implicit feedback in our study. We assume that this *consumption percentage* can serve as a behavioural measure of satisfaction with the video: when a user watches an entire video, this may be an indication of appreciation of the content. On the other hand, incomplete viewing could mean that the user does not like the content. In this experiment, the users were clearly instructed that they could stop the video if it did not match their interests.

An important remark should be made in this respect: especially in a real-life and non-controlled setting, users may have several reasons for stopping a video prematurely. Users may, e.g., be interrupted while watching or may prefer to watch the video at a more convenient time. These actions are not necessarily related to the video itself or do not necessarily say something about the user's appreciation. In order to not bias the viewing behaviour in the experiment (as investigated in Section 2.3.5), the consumption percentage was not taken into account by the recommender during the experiment.

2.2.2.3 Recruiting test subjects

Using a convenience panel-sampling method, a panel of test users was composed at the beginning of our study. Convenience sampling draws on easily accessible test subjects (e.g., university students) or people who volunteer to participate. For this experiment, our own Facebook network and that of our colleagues was used as a starting point for the recruitment of potential test users (by announcing the experiment in our status updates, sending personal messages, etc.). There are a

number of disadvantages to this type of sampling, the main ones being that the sample is not representative of the entire population and that people who volunteer to participate may be biased. Therefore, the results cannot be generalized. However, in this study, we believe that the research objectives justify the use of the convenience sampling method: the analyses of Section 2.3.3, 2.3.4, and 2.3.5 are based on individual video sessions and do not consider differences between the test users.

During the recruitment phase, we explained to potential test subjects that the experiment would consist of four successive 'waves' and that, with each wave, a questionnaire related to the (use of the) PersonalTV service would be sent out. We also mentioned that people who successfully participated in all four waves would receive a reward (i.e., a scratch card) and that they would stand a chance to win a gift check worth 25 euros.

2.2.2.4 Evaluation procedure

This study combines subjective, explicit user feedback, i.e., self-report data, with implicit user feedback, i.e., measured objective data related to the use of the application. The explicit feedback is gathered by means of the star-rating mechanism in PersonalTV and by online questionnaires. The implicit user feedback refers to the consumption percentage, i.e., the percentage of a video that is watched by a user and that is logged by the application. These data were gathered during four successive research waves, in which the test subjects were involved.

During each of the four research waves, a link to a structured online questionnaire was sent out. The test subjects were given sufficient time to complete the questionnaires: there was one week between every wave of the study during which the participants were given the opportunity to undertake the test at a convenient time. When necessary, reminders were sent out.

The first and most broad questionnaire was distributed in week one. It included topics such as the current use of Facebook and online video sites, the use of, interest in, and attitude towards existing video rating and recommender systems. In addition, a set of socio-demographical questions was included.

In the second questionnaire, the test subjects were introduced to the PersonalTV application. They were asked to log in on Facebook, to navigate to the application, and to select and watch three videos (on a desktop computer) according to their personal preferences and interests, using the 'most viewed', 'top rated', or 'keyword-search' features of the application. To minimize the burden on the test subjects, the number of videos they had to watch was limited. The instructions clearly mentioned that the test subjects could stop a video prematurely if it did not match their preferences. After each video watching, the test subjects had to evaluate the video using the star-rating mechanism of the PersonalTV application. In addition, a set of questions was asked via the online questionnaire in order to

gather explicit feedback immediately after each video. These questions dealt with the retrieval and selection of the video (e.g., Which aspects were important in the selection process? How was the video found? How satisfied or dissatisfied was the respondent with aspects such as the content of the video, the duration of the video, the description, image and sound quality...?). At the same time, a number of objective parameters related to the video itself and to the viewing and rating behaviour of the respondents were logged. More concretely, we logged the title, content category, tags, name of the uploader, view count on YouTube, mean rating on YouTube, and duration of the video as well as the percentage of the video that was watched by the user, the start time, and the user's rating.

In the third questionnaire, which was similar to the second one, the test subjects were asked to watch three videos and to answer the same set of questions (cfr. questionnaire two) after every video. Again, their viewing and rating behaviour was monitored. These data, in addition to the data captured in week two, enabled the recommendation algorithm to develop and refine the PersonalTV profile of every test subject based on the personal preferences inferred from his/her rating behaviour.

Finally, the fourth and last questionnaire again invited the test subjects to watch three videos and answer a set of questions (cfr. second and third questionnaire). However, this time, the test subjects were asked to select videos from their personal suggestions generated by the recommendation algorithm based on their PersonalTV profile. In this questionnaire, a number of additional questions were asked. For example, we wanted to know whether the recommended videos matched the interests and preferences of the test subjects (according to themselves). Moreover, we also included a set of questions on the application itself (e.g., Would you use it again? Do you like the current rating system? Do you intend to use the application in the future: yes or no (and why)? Which changes might improve the application?).

After the experiment, the data were checked in terms of their completeness and the answers to the questionnaires had to be linked to the explicit feedback data from the star-ratings, and the implicit feedback data from the viewing behaviour, which are logged in the PersonalTV storage system.

2.2.3 Sample description

The recruitment phase (as described in Section 2.2.2.3) resulted in a sample of 76 test subjects who expressed interest in the study and who participated in the first 'wave'. Of these 76 test subjects who started, 69 reached the end of the first questionnaire.

However, a considerable number of test subjects dropped out of the experiment at each successive wave. During the second, third, and fourth waves, 72, 46, and 42 test subjects, respectively, participated, meaning that from the original panel

of 76 test subjects, 42 people participated in all four waves. Finally, 37 of them reached the end of the last questionnaire. Although no formal drop-out analysis was performed, we believe that most of the drop-out was due to the timing of the experiment, which largely took place during the summer holidays. Moreover, the lack of motivation to answer the questionnaire and to watch the requested number of videos also played a role.

The original sample of 69 test subjects who reached the end of the first questionnaire consisted of 58.0% men and 42.0% women. The mean age of the participants in wave one was 26.9 years old (with a standard deviation of 6.6); the youngest test subject was 17 and the oldest, 56 years old. However, the panel size was reduced to 42 (of which 37 reached the end of the last questionnaire) in the last wave of the study. Therefore, the description of the socio-demographical profile of the test subjects in the remainder of this section is based on this final sample of test subjects.

An important remark should be made here: a somewhat surprising and interesting finding was that for a small minority of test subjects who had successfully completed all the questionnaires, no entries were found in the objective monitoring data. Two test subjects answered the questions related to nine videos, but neither watched a single video according to the logged data in the PersonalTV service. A possible explanation for the discrepancy between the objective and subjective data for some test subjects is that these participants were 'treasure hunters' who skipped the video watching and randomly filled in the questionnaire to receive the reward. Therefore, the data of these two test subjects were not used in the analysis. For some participants, the responses to the questionnaires (explicit feedback) did not completely correspond to the viewing behaviour (implicit feedback). But since the analysis of Section 2.3 was performed at the level of individual, watched videos using only data from test subjects who had completed all four questionnaires, the partial discrepancy between the objective and subjective data for some participants did not affect the results. If the self-reports had been the only form of user evaluation in this study, these anomalies would not be detected and the data would have been included in the analyses.

Of the final panel, 36.6% is female and 63.4% is male. Their ages range between 17 and 56, with an average of 27.0 years and a standard deviation of 6.5. When we take a look at the profession of the test subjects, we distinguish 61.0% as employees, 22.0% as students, 7.3% as blue-collar workers, 4.9% as executives, 2.4% with a free profession, and 2.4% as pensioners. The educational level (highest obtained degree) of the panel members varies between a master's degree (58.5%), academic bachelor's degree (12.2%), secondary school degree (9.8%), professional bachelor's degree (7.3%), and post-academic degree or PhD degree (4.8%). Another 7.4% have a degree that is different to those mentioned. A look at the family situation of the participants shows that about half of them are married

or living together (with or without children). Also, 29.3% are living with parents or family and 19.5% are living on his or her own.

2.3 Results

First, we discuss the results of the questionnaire of the first wave, dealing with the participants' current use of online video delivery systems. Next, we briefly introduce the measures that were used in the analysis of the data obtained by the second, third, and fourth wave. Then, we discuss the relation between the content retrieval method and the consumption percentage, between the content retrieval method and the reported satisfaction, and between the consumption percentage and the reported satisfaction. Finally, we give an overview of the collected general feedback from the test subjects.

2.3.1 Test subjects' current use of online video services

The PersonalTV client is implemented as a Facebook application, so the test subjects were firstly asked a number of general questions related to their use of Facebook. Of the 42 users who participated in each of the four questionnaires, 36 test subjects (85.7%) claimed to have used Facebook actively during the past week. On average, they had used Facebook four times per day and had spent 34 minutes on Facebook a day during that week. In the weekend, the reported daily average use of Facebook was 26 minutes and around three visits per day.

Furthermore, the respondents were asked whether they had used YouTube (e.g., for watching a video) during the past week. This was the case for 27 respondents (64.3%). On average, 13 minutes were spent on YouTube on a weekday (versus 10 minutes during an average weekend day). A number of other platforms similar to YouTube were also used by the panel members during the past month: Vimeo was used by 14%, Google Videos[7] by 21%, and the Flemish website GarageTV[8] by 21% of the test subjects.

We also presented the test subjects a number of social network features / activities (such as posting a photo or video, commenting on it, or giving a rating) related to online videos and asked them how frequently they had used these features during the past month (according to their own appraisal). Around 57% of the panel members mentioned having watched online video(s) at least once a week. In contrast, 81% never post a video online. Evaluating a video either by giving it a

[7]Google Videos (originally Google Video) was a free video sharing website, from Google Inc. In 2006, Google bought former competitor YouTube and in 2009, Google discontinued the ability to upload videos to Google's web servers. On August 20, 2012, Google Videos was shut down and the remaining Google Videos content was moved to YouTube.

[8]GarageTV is the Flemish alternative of YouTube developped and hosted by Telenet. Nowadays this video delivery system is called Zita.

rating or a comment is something that 73.2% and 56.1%, respectively, of the participants never do. Receiving videos from friends and sending videos to friends, on the other hand, is rather popular (only 12.3% of the participants has never done).

2.3.2 Measures

For the data analysis of Section 2.3.3, 2.3.4, and 2.3.5, three important measures are used:

- *Content retrieval.* For each video in waves two and three, the test subjects were asked whether they selected the video they watched either through YouTube's (a) most viewed videos, (b) top rated videos, (c) search engine, or (d) or some other way. In the fourth wave, respondents were obliged to use the algorithm's suggestions.

- *Objective measures.* The application recorded the percentage per video that was actually watched by the respondents (mean = 77.0, standard deviation = 32.9). For the analysis of Section 2.3.5, consumption percentage, being a behavioural measure of satisfaction with a video, was dichotomized using an arbitrary cut-off value of 90% (0-90% = incomplete consumption, N = 65 and 91-100% = (nearly) complete consumption, N = 106).

- *Subjective measures.* Firstly, a single Likert statement ranging from 1. absolutely not satisfied, 2. not satisfied, 3. satisfied nor dissatisfied, 4. satisfied, to 5. absolutely satisfied, measured satisfaction with a particular video's content (*satisfaction content*). Secondly, satisfaction with the way of content retrieval was likewise measured (*satisfaction retrieval*). Thirdly, satisfaction with the video's audiovisual quality was measured in a similar way by three Likert statements regarding image quality, sound quality, and their respective synchronization. These three items load on an internally consistent single factor ($\alpha = 0.74$), explaining 66% of the cumulative variance. Further, we refer to this combined variable as the *satisfaction quality* variable. Table 2.1 lists the subset of the questions asked in wave two, three, and four that were used to obtain these subjective measures.

2.3.3 The relation between the content retrieval method and the consumption percentage

We wanted to see if the way people choose videos (content retrieval) through either *most viewed*, *top rated*, *search*, or *PersonalTV suggestions* has an effect on how much of that video they watch (objective measure). In other words: does the

Reference	Question	Possible answers
1. Satisfaction Content	To what extent are you satisfied or dissatisfied with the content of the video?	5-point Likert statement
2. Satisfaction Retrieval	To what extent are you satisfied or dissatisfied with the way you found the video?	5-point Likert statement
3A. Satisfaction Quality	To what extent are you satisfied or dissatisfied with the image quality of the video?	5-point Likert statement
3B. Satisfaction Quality	To what extent are you satisfied or dissatisfied with the sound quality of the video?	5-point Likert statement
3C. Satisfaction Quality	To what extent are you satisfied or dissatisfied with the synchronization of image and sound in the video (by this we mean that image and sound are attuned)?	5-point Likert statement

Table 2.1: The subset of questions that were used to obtain the subjective measures used in the data analysis, together with a reference to these questions and the possible answers

content retrieval method affect the *consumption percentage*? To test for this potential effect of the way of content retrieval on the rates of actual consumption, an ANalysis of COVAriance (ANCOVA) was used. An ANCOVA evaluates whether population means of a dependent variable are equal across levels of a categorical independent or factor variable, while statistically controlling for the effects of other continuous variables that are not of primary interest, known as covariates [3]. The resulting F-test statistic stands for the ratio of the between-group variability (i.e., the explained variance) and the within-group variability (i.e., the unexplained variance) [4].

For this analysis, an ANCOVA was computed using the *content retrieval* type as a factor and *consumption percentage* as a dependent. As some of the test subjects had tried the application before the research took place, the exact number of *previous views* per test subject, as recorded by the application, was entered in the model as a covariate.

The results depicted in Figure 2.4 show that the means of *consumption percentage* do not differ significantly between the four *content retrieval* types ($F(3, 166) = 1.18$, $p > 0.05$); nor was *previous views* significantly correlated to *consumption percentage* ($F(1, 166) = 3.73$, $p > 0.05$). In summary, it appears that *content retrieval* has no effect on *consumption percentage* (objective data) and, thus, does

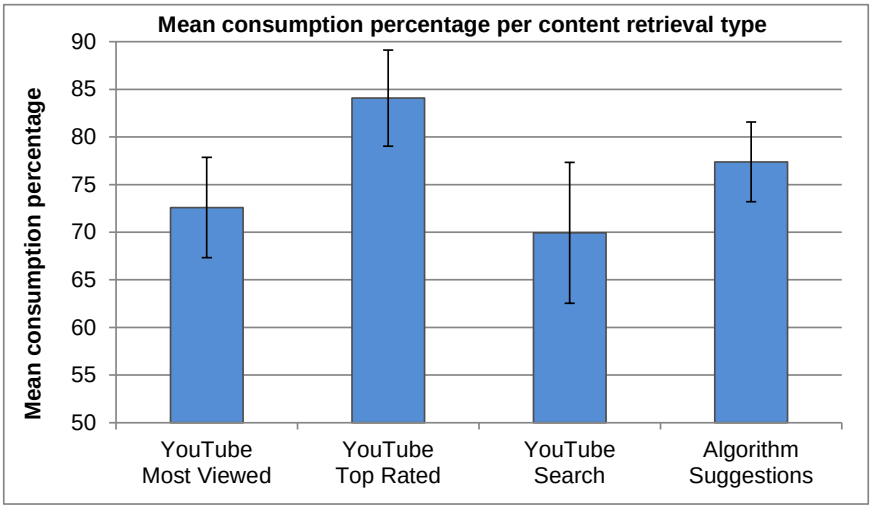


Figure 2.4: Mean consumption percentage per content retrieval type with the 95% confidence intervals

not allow for a prediction of whether the participant would watch the whole video or not.

2.3.4 The relation between the content retrieval method and the reported satisfaction

We also investigated if the content retrieval method has an influence on the reported satisfaction of the participants with the content of the video (*satisfaction content*), with its audiovisual quality (*satisfaction quality*) and with the way the video was chosen (*satisfaction retrieval*). In other words: does the *content retrieval* method affect the reported *satisfaction*? To investigate the influence of the content retrieval method on the reported satisfaction, a Multivariate ANalysis of COVAriance (MANCOVA) was used. A MANCOVA is an extension of ANCOVA methods to handle cases where there is more than one dependent variable and where the control of concomitant continuous independent variables - covariates - is required [5]. The Wilks' λ test statistic, which is a commonly used multivariate version of the ANOVA F-statistic, represents the ratio between the error variance (or covariance) and the effect variance (or covariance) [5].

For this analysis, a one-way MANCOVA was computed using the *content retrieval* type as a factor and the three subjective measures of *satisfaction* as dependents. The indication 'one way' in the name indicates that the analysis includes only one independent variable. As in the previous analysis, the exact number of *previous views* per test subject, was included as a covariate.

Although the MANCOVA analysis showed no significant differences between the subjective measures of satisfaction according to the content retrieval type, ($F(3, 167) = 31.26$, $p > 0.05$; $Wilks'\lambda = 0.93$), a number of trends can be observed. Firstly, as depicted in Figure 2.5, videos that are looked for and found using YouTube's search engine, have high values in terms of appreciation: if you watch a video that you have searched for on YouTube, it is likely that you will appreciate the content of the video. Secondly, videos that are most viewed (*YouTube most viewed*) tend to have a high audiovisual quality: they are most satisfying in terms of quality perception. Finally, selecting from *YouTube top rated* videos is preferred for content retrieval. In summary, although a number of trends can be identified, there are no significant differences in the reported *satisfaction* between the different *content retrieval* types.

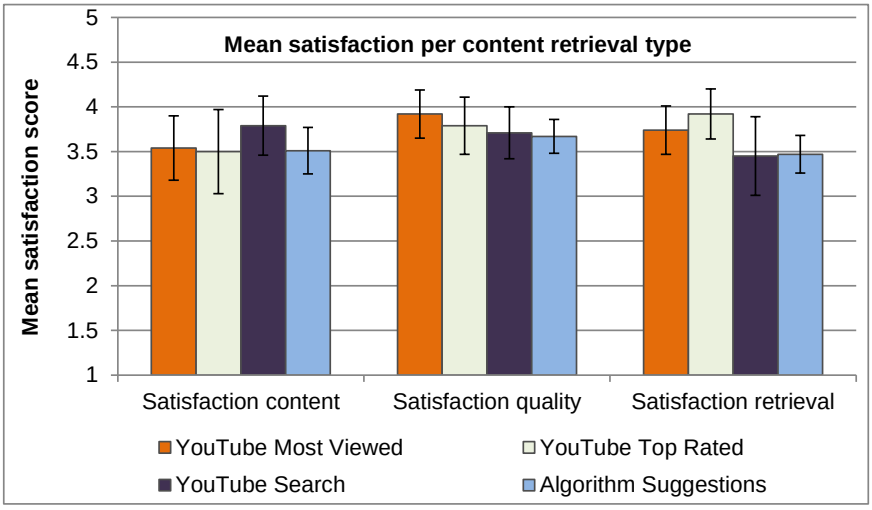


Figure 2.5: Mean satisfaction per content retrieval type with the 95% confidence intervals

2.3.5 The relation between the consumption percentage and the reported satisfaction

Finally, we also wanted to see if the *consumption percentage* (objective data) and the reported *satisfaction* (subjective data) converge. This comes down to the question: do implicit and explicit feedback converge? In our study this would imply that the percentage of a video that is watched by the user, is an indication of his or her (dis)satisfaction with the content, audiovisual quality, and/or content retrieval type. In order to answer this question, a Multivariate ANalysis Of VAriance (MANOVA) was used. MANOVA is a statistical test procedure for com-

paring multivariate (population) means of several groups. MANOVA is used when there are two or more dependent variables, and is therefore a generalized form of univariate analysis of variance (ANOVA) [4]. An ANOVA (analysis of variance) is used to determine whether or not the means of several groups are all equal, and therefore generalizes the T-test, (a statistical hypothesis test which compares the mean values of two groups,) to more than two groups [4].

To investigate the convergence of implicit and explicit feedback, a one-way MANOVA was computed using *consumption percentage* as a factor and the subjective measures of satisfaction (*satisfaction content*, *satisfaction quality*, and *satisfaction retrieval*) as dependents. More concretely, we used a dichotomized measure of consumption percentage with a cut-off value of 90%: incomplete consumption means that 0-90% of the video was watched, and (nearly) complete consumption means that 91-100% of the video was watched.

The MANOVA analysis showed a significant effect of *(in)complete consumption* on the combined dependent *satisfaction* variable, ($F(3, 167) = 3.92$, $p < 0.05$; $Wilks'\lambda = 0.93$). Although an univariate ANOVA analysis indicates significant effects of the factor on all three dependent variables ($p < 0.05$), further analysis using a Bonferroni adjusted p-level (of $0.05/3 = 0.017$) suggests that there is no significant effect on the satisfaction with audiovisual quality (*satisfaction quality*), ($F(1, 169) = 3.80, 0.017 < p = 0.02 < 0.05$), and on the satisfaction with the way of content retrieval (*satisfaction retrieval*), ($F(1, 169) = 4.53, 0.017 < p = 0.035 < 0.05$). Put differently, the way that the video is found does not influence whether it will be watched completely or incompletely. The same goes for the audiovisual quality, so, e.g., even when the technical quality of a video is poor, the user may still finish watching.

On the other hand, there is a significant effect of *consumption percentage* on the satisfaction with the content, ($F(1, 169) = 7.86$, $p = 0.006 < 0.0017 < 0.05$). This indicates that the difference between incomplete viewing and (nearly) complete viewing for *satisfaction content* in Figure 2.6 is substantial. Respondents who watch more than 90% of a video (objective data) report a significantly higher satisfaction with the video's content (subjective data), suggesting a convergence of both data sources. In other words, people tend to watch the (nearly) entire video if they like it and if they watch only a part of the video, it is likely that they are less happy with the content.

2.3.6 Qualitative feedback from the test subjects

We also asked the respondents whether they intend to continue using the application or not, and equally important in case of the latter: why not? Although 48.6% intended to keep using the application, 51.4% of the test subjects responded

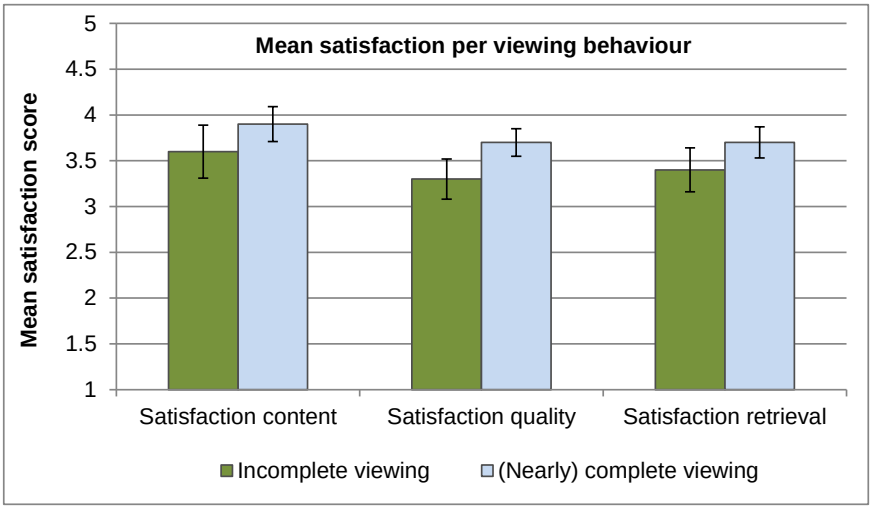


Figure 2.6: Mean satisfaction per viewing behaviour with the 95% confidence intervals

negatively to this question. An analysis of the responses to the 'why?' question indicates a number of reasons.

A number of test subjects believe that there are already sufficient platforms for online video watching and sharing with various possibilities and affordances[9]. They find no substantial added value in PersonalTV compared to sites such as YouTube or Vimeo, which offer the same content, similar recommendation services, and more features for retrieval. In this respect, one of the respondents mentioned the 'queue' and 'add to playlist' options on YouTube: *"Videos can't be 'marked' for a future viewing session like on YouTube. This would, however, make it a lot easier to retrieve them without too much effort."* The lack of compatibility with existing online video platforms is also mentioned. It is argued that a better integration might make the application more attractive for users: *"If the application would be built in a site like YouTube, then I would probably use the application more often. Now, you need to make a detour by logging in to Facebook to start the application."*

This procedure, which requires users to log in on Facebook to use the application and receive personalized suggestions, is experienced as time-consuming and inhibiting by several test subjects. On YouTube, e.g., even non-registered users receive recommendations based on their previous viewing behaviour. The detour via Facebook, thus, requires an additional effort that some test subjects are not willing to make.

[9] An affordance is a quality of an object, or an environment, which allows an individual to perform an action.

Addressing the key asset of the application, a number of test subjects mentioned that the recommendations that they received were not sufficiently refined and, therefore, not matching their particular interests. They state that there is still a significant margin for improvement. An important factor in this respect could be the fact that for most of the participants, the number of evaluated videos is still rather limited (cfr. cold start problem) and the videos are possibly very diverse in terms of content. The cold start problem is a typical problem of recommender systems. As recommendation algorithms suggest items based on users' past preferences, new users have to rate a sufficient number of items to enable the system to capture their preferences accurately and thus provide reliable recommendations. As a result, it is (more) difficult for the PersonalTV recommendation algorithm to make good suggestions. In this respect, one person argues, *"Since these recommendations are an aggregation of previous behaviour and since the offer is always a common denominator, I don't intend to use the application again. I have more faith in a human editor (e.g., a friend) for finding and suggesting interesting content for me."*

For most other participants, the main reason(s) for having no intention to use the application in the future is/are the lack of time and lack of interest in watching online videos: *"I rarely watch videos online and never proactively look for them. I don't feel the need to spend time on this."* A final important reason that was given for not adopting the PersonalTV application is that of 'habit'. As one test subject phrases it, *"For me it is a matter of habit. I always go to YouTube when I want to watch online video content. I am used to the interface as well."*

At the level of suggestions for improvement, it was mentioned that a more expanded content offer (e.g., premium content) might help to improve the user experience. Also, the possibility to change settings related to the suggestions might be useful: one of the respondents mentions that *"it would be better if more suggestions could be offered and if these could be better sorted by category; for example, I'm interested in comedy and music, but that doesn't mean that I'm always in the mood to watch both."* Another suggestion concerns the ability to specify a number of personal preferences in the profile so that the user can also build up his or her profile by indicating, e.g., preferred genres, the fact that he or she only watches short videos, etc.

Finally, some usability and Quality of Service (QoS) issues were identified as well. In terms of usability, it was mentioned that it should be easier and clearer to navigate through the list of thumbnails and to scroll down. Since several participants made this remark, it should be more closely investigated and changed if necessary. One test subject also had difficulties due to an incompatible screen resolution. In terms of quality of service, one participant experienced problems with the web server, which could not establish a connection to YouTube. Since there was only one participant who reported on this issue, but the problem was encoun-

tered several times, it may be due to the blocking of YouTube on the computer or the network of the participant.

2.4 Conclusions

This chapter presented results from a study aimed to evaluate a recommender system from a user point of view, thereby providing insights into the users' interaction behaviour and experience with the system. To this end, explicit, subjective user feedback was complemented with implicit, objectively gathered data. By involving a panel of test subjects, we tried to gather insights that might help to optimize and refine recommender systems. To enable the logging of different objective parameters, the PersonalTV service was developed and used as test system. The PersonalTV service enables its users to watch YouTube videos and builds up a personal viewing profile in order to give personalized content suggestions.

We investigated whether the content retrieval method (YouTube most viewed, YouTube top rated, YouTube search, or algorithm suggestions) had an influence on the consumption percentage (objective data) and on the reported satisfaction (subjective data) of the test subjects. In this respect, we found that the *content retrieval* method had no significant effect on the *consumption percentage*. A similar conclusion holds for the reported satisfaction of the test subjects. Although a number of trends were identified, the four *content retrieval* types do not yield significant differences in terms of *user satisfaction* with the content, audiovisual quality, and way of content retrieval.

We also investigated whether the measured objective user interaction (implicit feedback) and the subjective evaluations (explicit feedback) converge. The results indicate that there is no significant correlation between the objective measure *consumption percentage* and two of the satisfaction measures (*satisfaction quality* and *satisfaction retrieval*). In contrast, the *consumption percentage* has a significant influence on the *satisfaction with the content*: (nearly) complete viewing yields a significantly higher satisfaction with the content of the video thus suggesting a convergence of both measures and implying that consumption percentage could be used as an indirect rating mechanism. This implicit feedback mechanism might help to lower the burden on the user of having to rate every watched video and it might lessen the dependence of the algorithm on explicit user feedback.

An important remark should be made in this respect: although we assumed that playing a video until the end or stopping a video prematurely could provide a cue on the user's appreciation of this video, this assumption may not always hold true and should, thus, be used with care. Users may have other reasons for stopping a video before the end: e.g., the viewing session is interrupted, the user wants to watch the video at a more convenient time, etc.; and these behavioural actions do not necessarily represent an implicit evaluation of the (recommended) content in

terms of approval or disapproval. The correct interpretation of this implicit feedback is especially challenging in a natural viewing context in which there is no control over the viewing conditions. It is also crucial, since incorrectly interpreted implicit feedback may affect the accuracy of the recommendations that are made to the users in a negative way. Some of the issues related to gathering and interpreting of implicit feedback could be addressed in future research by analysing the viewing sessions (and related monitored data) at a higher level. By combining video metadata with timestamps and other logged viewing-related data, it could, e.g., be investigated whether a video was watched within a consumption spurt (i.e., a sequence of viewed videos) or at the end of a spurt.

Evaluating the PersonalTV service as a recommender system use case provided insights into the users' expectations and experiences. Although some of the suggestions and comments made by the panel members relate to more general issues such as the login via Facebook, the lack of certain features for video retrieval and storage, the fact that the user cannot control his or her profile or make it more complete by specifying personal preferences, etc., these aspects affected the participants' experience and satisfaction with the application. A better integration of relevant personal information and of preferences specified by users themselves in the recommendation algorithm, can help to increase the accuracy of the recommendations and to reduce the time that is needed to build up a user profile.

The collected feedback helps to understand drawbacks and barriers to use, reuse, or prolonged use of a recommendation application. In view of the evaluation of a recommendation application, this temporal aspect is of vital importance since the accuracy of the suggestions largely depends on the completeness of the user's profile and, thus, on the frequent use of the application.

References

[1] W. A. Warr. *Social software: fun and games, or business tools?* Journal of Information Science, 34(4):591–604, August 2008.

[2] G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Co, New York, NY, USA, 1983.

[3] A. Rutherford. *Introducing ANOVA and ANCOVA: a GLM approach.* Sage Publications Limited, 2001.

[4] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models.* McGraw-Hill, fifth edition, 2005.

[5] G. Garson. *GLM: MANOVA and MANCOVA.* Statnotes: Topics in multivariate analysis, 2005.

3

User-centric evaluation of recommendation algorithms

3.1 Introduction

In this chapter, the focus is on evaluating recommendation algorithms from a user point of view. For the comparison of algorithms, often offline metrics like the RMSE, MAE, or precision and recall are calculated [1], as explained in chapter 1. These kinds of metrics allow automated and objective comparison of the accuracy of the algorithms but they alone cannot guarantee user satisfaction in the end [2]. As shown in [3], the use of different offline metrics may even lead to a different outcome of the 'best' algorithm for the job. What is more, Hayes et al. [4] state that real user satisfaction can only be measured in an online context.

In this chapter, the satisfaction for real-life users is assessed by means of an online user-centric evaluation. Five different recommendation algorithms are compared in terms of accuracy, familiarity, novelty, diversity, transparency, satisfaction, trust, and usefulness. Furthermore, the relationship between these qualitative attributes is investigated.

This online evaluation was performed in the context of recommending cultural events for the users of 'Uit in Vlaanderen'[1], a Belgian cultural events website. This website contains the details of more than 30,000 near future and ongoing

[1] http://www.uitinvlaanderen.be/

cultural activities including movie releases, theatre shows, exhibitions, fairs, and many others[2].

In the research domain of recommender systems, numerous studies have focused on recommending movies. They have been studied thoroughly and many best practices are known. The area of event recommendations on the other hand is relatively new. Events are so called *one-and-only* items [5], which makes them harder to recommend. Whereas other types of items generally remain available (and thus recommendable) for longer periods of time, this is not the case for events. They take place at a specific moment in time and place to become irrelevant very quickly afterwards.

Some approaches towards event recommendation do exist. For the Pittsburgh area, a cultural event recommender was built around trust relations [6]. Friends could be explicitly and implicitly rated in terms of trust ranging from 'trust strongly' to 'block'. A recommender system for academic events [7] focused more on social network analysis in combination with Collaborative Filtering (CF) and finally, a hybrid event recommendation algorithm [8] was described as an approach in which both aspects of CF and Content-Based (CB) algorithms were employed. To our knowledge, event recommendation algorithms were never compared in a user-centric designed experiment with a focus on optimal user satisfaction.

3.2 Test setup

3.2.1 Goals of the study

The end goal of this research is to improve the user satisfaction for real-life visitors of event websites. For that reason, an online user-centric evaluation experiment is performed to compare commonly-used recommendation approaches. Whereas several studies on event recommendations already exist in literature, such a comparison study including different algorithms remains absent. So, the goal of this experiment is, in the short-term, to discover the 'best' recommendation approach for event websites according to the users, and in the long-term, to improve the user satisfaction of these recommendation services.

Both explicit and implicit feedback in the form of user interactions with the website were logged over a period of 41 days, serving as the input for the five recommendations algorithms used in this experiment. By means of a questionnaire, test subjects were asked to rate different qualitative aspects of the recommender system, not only accuracy, but also more indirect characteristics such as user satisfaction and trust. These subjective evaluations can reveal the correlations and any causal relationships between the qualitative aspects. The results of this experiment

[2]Our work regarding personal recommendations for events was in cooperation with the MultiMedia Lab (MMLab) research group, Department of Electronics and Information Systems, Ghent University.

can be used to choose the most optimal recommendation approach, depending on the qualitative aspects that have to be optimized.

3.2.2 Procedure

In this section, we elaborate on the specifics of the experiment such as the feedback collection, the recommendation algorithms, how we randomized the users, and the questionnaire.

3.2.2.1 Gathering feedback

Feedback collection is a very important aspect of the recommendation process. Since the final recommendations can only be as good as the quality of their input, collecting as much high quality feedback as possible is of paramount importance. Previous feedback experiments we ran on the website [9] showed that collecting explicit feedback (in the form of explicit ratings) is very hard, since users do not rate often. Clicking and browsing through the event information pages are on the other hand activities that were abundantly logged. Now, various activities on the event information page actually indicate a user preference for the event. E.g., clicking the 'share on Facebook' or 'share on Twitter' button, querying the itinerary, printing the event's information, querying information regarding public transportation to reach the event, querying additional dates and locations or details about the event, mailing the event to a friend, or just browsing the event information page.

Since explicit ratings are typically provided after an event has been visited, and feedback is usually gathered when the event is not available for attendance any more, CF algorithms would be useless if they are only based on these explicit ratings. It therefore makes sense to utilize also implicit feedback indicators such as printing the event's information, which can be collected before the event has taken place. For optimal results, we therefore monitored explicit feedback in the form of clicks on 'I like this' as well as implicit feedback in the form of user interactions with the event website. In total 11 distinct feedback activities were combined into a feedback value that expresses the interest of a user for a specific event.

The different activities are listed in Table 3.1 together with their resulting feedback values which were intuitively determined. These feedback values range from 1 if we are absolutely sure that a user is fond of the event (e.g., if the user explicitly stated that he/she likes the event), down to 0.3 if we are far less sure (e.g., if the user only browsed the event information page). The $max()$ function is used to aggregate multiple feedback values in case a user provided feedback in more than one way for the same event. E.g., if the user only browsed the event page, a feedback value of 0.3 is registered, but if the user also queries additional informa-

tion about public transportation, a feedback value of 0.6 is registered. (Alternative functions such as the $sum()$ are of course possible.)

Feedback activity	Feedback value
Click on 'I like this'	1.0
Share on Facebook/Twitter	0.9
Click on Itinerary	0.6
Click on Print	0.6
Click on 'Go by bus/train'	0.6
Click on 'Show more details'	0.5
Click on 'Show more dates and locations'	0.5
Mail to a friend	0.4
Browse to an event page	0.3

Table 3.1: The activities that were logged as user feedback together with the feedback value indicating the interest of an individual user for a specific event

3.2.2.2 Recommendation algorithms

The collected feedback is used as input for the recommender system to generate personalized suggestions for the test subjects. In this experiment, five different algorithms are used, each of which generated a list of personal recommendations for every test subject. Each test subject, unaware of the different algorithms, is randomly assigned to one of the five user groups receiving recommendations generated by one of these algorithms as described in Section 3.2.2.3. For each algorithm, the final event recommendations are checked for their availability and familiarity with the user. Events that are not available for attendance any more (availability), or events that the user has already explored by viewing the webpage, or clicking the link (familiarity), are replaced in the recommendation list.

Bollen et al. [10] hypothesized that a set of somewhere between seven and ten recommended items would be ideal in the sense that it can be quite a diverse set but still manageable for the users. Therefore, the test subjects of this experiment received a recommendation list containing eight events together with an online questionnaire to evaluate different qualitative aspects of their recommendations, as discussed in Section 3.2.2.4.

As explained earlier, the focus of this research is not on developing a new recommender but rather on investigating the qualitative aspects of existing recommendation approaches in the context of event recommendation. Therefore, a number of state-of-the-art recommendation algorithms are used: a CB recommendation algorithm, a Nearest Neighbourhood (NN) CF technique, a hybrid CF-CB

algorithm (Hybrid), and a recommendation algorithm based on Singular Value Decomposition (SVD). As a baseline recommendation algorithm, we used the random recommender (RAND).

3.2.2.2.1 Content-Based algorithm (CB) CB recommendation algorithms generate personalized recommendations based on the metadata of the content items. Given the ability of CB algorithms to recommend items before they received any feedback, they can perfectly handle the transiency of events. As a CB solution, the *InterestLMS predictor* of the open-source implementation of the Duine framework [11] is adopted (and extended to consider extra metadata attributes).

Based on the metadata attributes of the content items and the user's feedback for these items, the recommender builds a profile model for every user. This profile contains an estimation of the user's preference for each item and each metadata field that is linked to an item that received feedback from the user. Based on the preferences of this profile, the recommender predicts the user's preferences for new, unexplored items. Subsequently, the items with the highest prediction score are selected for the recommendation list.

The event metadata that was available for this experiment contains the title, the categories, the artist(s), and keywords originating from a textual description of the event. A weighting value is assigned to the various metadata fields (see Table 3.2), thereby attaching a relative importance to the fields during the matching process (e.g., a user preference for an artist is more important than a user preference for a keyword of the description). The employed keyword extraction mechanism is based on a Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme, and includes features as stemming and filtering stop words.

Metadata field	Weight
Artist	1.0
Category	0.7
Keyword	0.2

Table 3.2: The metadata fields used by the CB recommendation algorithm with their weights indicating their relative importance

3.2.2.2.2 Collaborative Filtering (CF) The used implementation of CF is based on the work of Breese et al. [12]. This nearest neighbour collaborative filter uses the Pearson correlation metric for discovering similar users or similar items based on the consumption behaviour (i.e., explicit or implicit feedback) of all users in the system.

In the user-based approach (UBCF), the user's preference for an item is predicted based on the preferences of similar users. The obtained prediction score estimates how much the item will be appreciated by the user. The items with the highest prediction score are included in the recommendation list for this user. In the item-based approach (IBCF), the user's preference for an item is predicted based on his/her preferences for similar items in the system. Again, the items with the highest prediction score are recommended to this user.

In this experiment, we opted for the user-based nearest neighbour approach (UBCF) because of the higher user-user overlap compared to the item-item overlap. Neighbours were defined as being users with a minimum overlap of 1 event in their feedback profiles but had to be at least 5% similar according to the Pearson similarity metric.

3.2.2.2.3 Hybrid recommender (Hybrid) The CF and CB recommender both have disadvantages: CB-only algorithms might produce recommendations with a limited diversity [2], and CF techniques might produce suboptimal results due to a large amount of unrated items (cold start problem). A hybrid recommendation algorithm, combining features of both CB and CF techniques, can (partially) eliminate these drawbacks.

The Hybrid recommender used in this experiment combines the recommendations with the highest prediction score of the CF and CB recommender into a new recommendation list. This algorithm acts on the resulting recommendation lists produced by the CF and CB recommender, and does not change the internal working of these individual algorithms. Both lists are interwoven while alternately switching their order such that both lists have their best recommendation on top in 50% of the cases. To avoid doubles, items that are recommended by the CF as well as by the CB recommender are only included once in the resulting list. The result is an alternating list of the best recommendations originating from the CF and CB recommender.

3.2.2.2.4 Singular Value Decomposition (SVD) Because of their excellent performance, recommendation algorithms based on matrix factorization are commonly used. Singular value decomposition (SVD) [13] is a well-known matrix factorization technique that addresses the problems of synonymy, polysemy, sparsity, and scalability for large data sets.

We opted for the open-source implementation of the SVD Recommender of the Apache Mahout project (version 0.6) [14] in this research. Based on preceding simulations on an offline data set with historical data of the website, the parameters of the algorithm were determined: 50 iterations were used to train the model and the number of features was set at 70.

3.2.2.2.5 Random recommender (Rand) To compare the results of the different recommenders, the random recommender was introduced as a baseline suggestion mechanism. This random recommender generates recommendations by performing a random sampling of the available events. The only requirement of these random recommendations is that the events are still available (i.e., it is still possible for the user to attend the event). The evaluation of these random recommendations allows to investigate if users can distinguish random events from personalized recommendations, and if so, the relative (accuracy) improvement of more intelligent algorithms over random recommendations.

3.2.2.3 Recruiting test subjects

For this experiment, test subjects were recruited from the existing users of the event website. These users were requested to participate in the experiment by an invitation in the newsletter and by a banner on the website. Users who were interested to participate were explained that their clicking behaviour on the website would be tracked in order to generate personal recommendations.

Test subjects who subscribed for the experiment, but who had not been active on the website after a period of 28 days, received a reminder e-mail to encourage them to use the website. The test subjects' activities on the website were monitored during a period of 41 days and the resulting data was used as input for the five algorithms to generate personal recommendations.

Since certain test subjects have provided only a limited amount of feedback during the experiment, not all algorithms were able to generate personal recommendations for these users. CF algorithms, for instance, can only identify neighbours for users who have overlapping feedback with other users (i.e., provided feedback on the same event as another user). Without these neighbours, CF algorithms are not able to produce recommendations.

Therefore, test subjects with a limited amount of feedback, hindering (some of) the algorithms to generate (enough) recommendations for that test subject, are treated separately in the analysis. Many of these test subjects were not very active on the website or did not finish the evaluation procedure as described in Section 3.2.2.4. This group of cold-start users received recommendations from a randomly assigned algorithm that was able to generate recommendations for that test subject based on the limited profile. Since the random recommender can produce suggestions even without user feedback, at least 1 algorithm was able to generate a recommendation list for every test subject. The comparative evaluation of the five algorithms however, is based on the remaining test subjects, who provided sufficient feedback for all algorithms. Each of these remaining test subjects is randomly assigned to one of the five algorithms, which generates the personal suggestions for that test subject. This way, the five algorithms, as described in Section 3.2.2.2, are evaluated by a number of randomly selected users of the website.

Subsequently, test subjects were informed via e-mail about the availability of these recommendations on the website. Herewith, they were asked to fill in a digital questionnaire to evaluate qualitative aspects of their recommendations, as described in Section 3.2.2.4. Again, a reminder e-mail was sent to encourage the test subjects if they had not yet completed the questionnaire five days after receiving the recommendations.

3.2.2.4 Evaluation procedure

While prediction accuracy of ratings used to be the only evaluation criteria for recommender systems, during recent years optimizing the user experience has increasingly gained interest in the evaluation procedure [2]. Existing research has proposed a set of criteria detailing the characteristics that constitute a satisfying and effective recommender system from the user's point of view. To combine these criteria into a more comprehensive model that can be used to evaluate the perceived qualities of recommender systems, Pu et al. have developed an evaluation framework for recommender systems [15]. This framework aims to assess the perceived qualities of recommenders such as their usefulness, usability, interface and interaction qualities, user satisfaction of the systems, and the influence of these qualities on users' behavioural intentions including their intention to tell their friends about the system, the purchase of the products recommended to them, and the return to the system in the future.

Therefore, we adopted (part of) this framework to measure users' subjective attitudes based on their experience towards the event recommender and the various algorithms tested during our experiment. Via an online questionnaire, test subjects were asked to answer 14 questions on 5-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (5) regarding aspects as recommendation accuracy, novelty, diversity, satisfaction, and trust in the system. As explained in chapter 1, some of these qualitative aspects, such as the user's trust in the system, are impossible to derive from offline evaluations. By means of an online questionnaire however, these subjective quality aspects can be assessed. Table 3.3 lists the 8 most relevant questions for this research, which are directly related to one of the quality aspects of the event recommender system. To check the consistency of the test subjects' answers, some of the questions were asked using a reverse scale.

As on many websites, there were no explanations for the recommendations in order to keep the recommendation block on the website as compact as possible. Therefore, the transparency aspect measures the extent to which the users were expecting the received recommendations based on their previous activities on the website, rather than the extent to which the recommendations can be clarified by using explanations.

Reference	Quality metric	Question
Q1	accuracy	The items recommended to me matched my interests.
Q2	familiarity	Some of the recommended items are familiar to me.
Q4	novelty	The recommender system helps me to discover new items.
Q5	diversity	The items recommended to me are similar to each other (reverse scale).
Q7	transparency	I didn't understand why the items were recommended to me (reverse scale).
Q8	satisfaction	Overall, I am satisfied with the recommender.
Q10	trust	The recommender can be trusted.
Q13	usefulness	I would attend some of the events recommended, given the opportunity.

Table 3.3: The questions that were used to evaluate the recommendations of the event website, together with a reference to these questions

3.2.3 Sample description

Almost 60000 registered users of the event website received an invitation for the experiment via the newsletter. In total 612 users responded positively to the request to participate in the experiment. So, we achieved an acceptance rate of around 1%, which is not abnormal for an online experiment. Of these 612 users who were interested, 232 actually completed the online questionnaire regarding their recommendations. After removal of questionable samples (e.g., users who answered every question with the same value, including the questions with a reverse scale) and users with insufficient feedback for the algorithms, 193 users remained. They had by average 22 consumptions (i.e., expressed feedback values for events) and 84% of them had 5 or more consumptions. The final distribution of the test subjects across the algorithms is displayed in Table 3.4.

Algorithm	Number of test subjects
CB	43
CB+UBCF	36
RAND	45
SVD	36
UBCF	33

Table 3.4: The five algorithms compared in the user-centric evaluation and the number of test subjects that actually completed the questionnaire about their recommendation list

3.3 Results

3.3.1 Subjective evaluations

Figure 3.1 shows the averaged results of the answers provided by the 193 test subjects in this experiment for the 8 questions we described in Section 3.2.2.4 and for each algorithm. These averaged results are used as an estimation of the quality of the algorithm regarding a specific aspect. The error bars indicate the 95% confidence intervals of these average values.

Evaluating the answers to the questionnaire showed that the hybrid recommender (Hybrid) achieved the best averaged results to all questions, except for question Q5, which asked the user to evaluate the similarity of the recommendations (i.e., diversity in reverse scale). For question Q5, the random recommender obtained the best results in terms of diversity, since random suggestions are rarely similar to each other.

The CF algorithm was the runner-up in the evaluation and achieved a second place after the hybrid recommender for almost all questions (again except for Q5, in which CF was the fourth after the random recommender, the hybrid recommender, and SVD).

The recommendations of the CB algorithm were moderately appreciated by the test subjects. For most questions, the averaged results of the CB recommender were worse than the results of the CF and Hybrid recommender but better than the SVD and random recommender. One exception is Q5, in which the CB recommender received the worst average score on diversity. Because the CB algorithm recommends the items that are most similar (in terms of metadata) to the items that the user has already consumed in the past, the test subjects evaluated these recommendations as the most similar to each other.

The average performance of SVD was lower than expected by achieving the worst results for questions Q1, Q7, Q8, Q10, Q13 and the second worst results

(after the random recommender) for questions Q2, and Q4. So surprisingly, the SVD algorithm performs (averagely) worse than the random method on some fundamental questions like for example Q8, which addresses the general user satisfaction. We note however that the difference in values between SVD and the Rand algorithm was not found to be statistically significant except for question Q5.

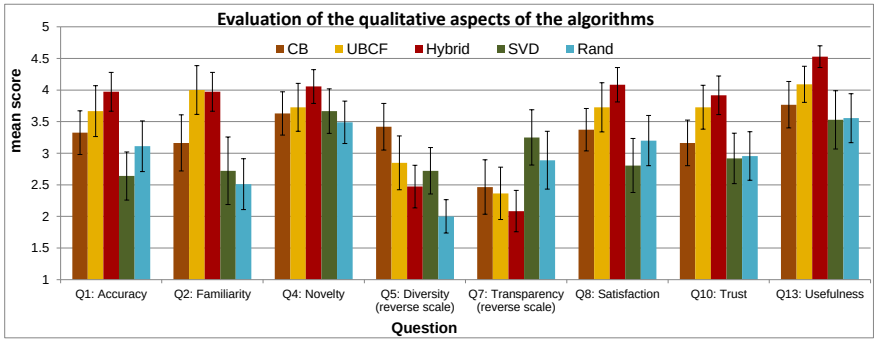
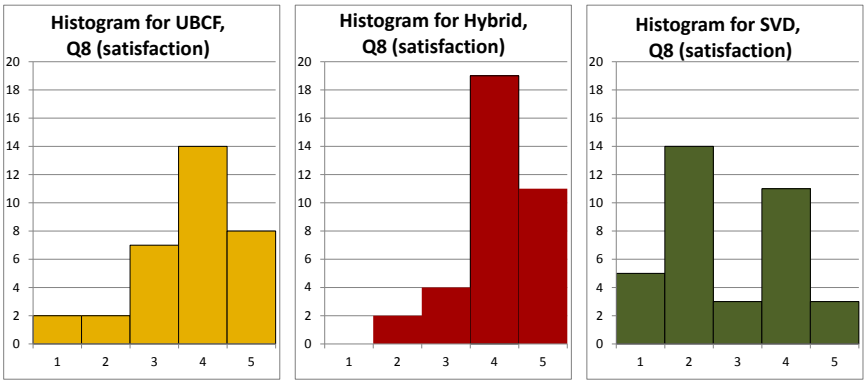


Figure 3.1: The averaged answers (on a 5-point Likert scale) of the evaluation questionnaire for each algorithm and the corresponding error bars indicating the 95% confidence intervals of the average values

This observation is further investigated by the histogram (Figure 3.2) of the different values (1 to 5) for the answers provided for question Q8. A clear distinction between the histogram of the SVD algorithm and the histograms of the other algorithms (UBCF and Hybrid shown in Figure 3.2) can be seen. Whereas for UBCF and Hybrid most values are grouped towards one side of the histogram (i.e., the higher values), this is not the case for SVD. It turns out that the opinions about the general satisfaction of the SVD algorithm where somewhat divided between good and bad with no apparent winning answer. Approximately half of the test subjects are dissatisfied with their suggestions, providing a rating of 1 or 2; and half of the test subjects are pleased, evaluating the suggestions with 4 or 5. These noteworthy rating values for the SVD recommender are not only visible in the results of Q8, but also for other questions like Q2 and Q5. These findings indicate that SVD works well for many users, but also provides inaccurate recommendations leading to dissatisfaction for a considerable number of other users. These inaccurate recommendations may be due to a limited amount of user feedback and therefore sketchy user profiles.

As expected, the random recommender obtained for most qualitative attributes a poor average score. Since these random recommendations are mostly unrelated to previously consumed content, most recommendations are not familiar to the test subjects and the random recommender obtained a low average score regarding



(a) UBCF (b) Hybrid (c) SVD

Figure 3.2: The histogram of the values (1 to 5) that were given to question Q8 (satisfaction) for the algorithms UBCF, Hybrid, and SVD

trust. In terms of diversity, the random recommender generates the most diverse recommendation list. Still a considerable number of test subjects thinks this list contains (too) similar recommendations (average rating of 2, including ratings of 4 and 5).

The success of the hybrid recommender is not only clearly visible when comparing the average scores for each question (Figure 3.1), but the hybrid recommender also showed to be statistically significantly better than every other algorithm (except for the CF recommender) for the majority of the quality metrics (accuracy (Q1), familiarity (Q2), satisfaction (Q8), trust (Q10), and usefulness (Q13)). A one-way ANalysis Of VAriance (ANOVA) relies on the restrictive assumptions of homogeneity of the variances of the distributions and normality of the distributions of the residuals [16]. Also the commonly-used T-test, a statistical hypothesis test which compares the mean values of two groups, relies on the assumption that the samples follow a normal distribution [16]. Since the test subjects' evaluations are discrete values, these assumptions may not apply. Therefore, the five recommendation algorithms were compared using the Wilcoxon rank-sum test as alternative. The Wilcoxon rank-sum test is a non-parametric statistical hypothesis test for assessing whether one of two samples of independent observations tends to have larger values than the other [17]. This way, the subjective evaluations of the qualitative aspects of the recommendations were compared according to a Wilcoxon rank-sum test using the recommendation algorithm as the grouping variable (independent variable). Table 3.5 shows the algorithms and qualitative aspects for which statistically significant differences ($p < 0.05$) could be noted according to this non-parametric statistical hypothesis test. Note that the matrix is symmetric.

	CB	UBCF	Hybrid	SVD	RAND
CB	-	familiarity, diversity, trust	accuracy, familiarity, diversity, satisfaction, trust, usefulness	accuracy, diversity, transparency, satisfaction	familiarity, diversity
UBCF	familiarity, diversity, trust	-	usefulness	accuracy, familiarity, transparency, satisfaction, trust	familiarity, diversity, trust
Hybrid	accuracy, familiarity, diversity, satisfaction, trust, usefulness	usefulness	-	accuracy, familiarity, transparency, satisfaction, trust, usefulness	accuracy, familiarity, novelty, diversity, transparency, satisfaction, trust, usefulness
SVD	accuracy, diversity, transparency, satisfaction	accuracy, familiarity, transparency, satisfaction, trust	accuracy, familiarity, transparency, satisfaction, trust, usefulness	-	diversity
Rand	familiarity, diversity	familiarity, diversity, trust	accuracy, familiarity, novelty, diversity, transparency, satisfaction, trust, usefulness	diversity	-

Table 3.5: The complete matrix of statistically significant differences between the algorithms on all the qualitative aspects using the Wilcoxon rank-sum test on a confidence level of 0.95.

3.3.2 Relating the quality aspects

Figure 3.1 seems to indicate that some of the answers to the questions are highly correlated. One clear example is question Q1, about whether or not the recommended items matched the user's interest, and question Q8, which asked about the general user satisfaction. As obvious as this correlation may be, other correlated questions may not be so easy to detect by inspecting the averaged results and so we calculated the complete correlation matrix for every question over all the algorithms using the Pearson correlation metric (Table 3.6). Due to the Pearson metric, values are distributed between -1.0 (negatively correlated) and 1.0 (positively correlated), and the matrix is symmetric. Note that questions Q5 and Q7 were in reverse scale. The bold values indicate the statistically significant correlations on a confidence level of 0.95.

From the correlation values, an interesting trend can be noted for the questions Q1, Q8, and Q10, dealing with respectively the accuracy of the recommendations, the user satisfaction, and user's trust of the system. The answers to the questions

regarding these three quality metrics are highly correlated to each other (very significant $p < 0.01$). Although correlation should not be confused with causality, the data indicate the strong relation between recommendation accuracy, user satisfaction, and trust of the system.

The strong correlation between transparency and the other quality metrics (such as satisfaction and trust), may be another reason why SVD performed worse than expected in the subjective evaluations of the experiment. Its inner workings are the most obscure and least obvious to the user and therefore also the least transparent. This limited transparency may have a negative influence on the user satisfaction and his/her trust in the system.

Another interesting observation lies in the correlation values of question Q5. The answers to this diversity question are almost completely unrelated to every other question (i.e., low correlation values which are not significant $p > 0.05$). It seems like the test subjects of the experiment did not value the diversity of a recommendation list as much as the other aspects of the recommendation system. The averaged results (Figure 3.1) of the answers on the diversity question (lower is more diverse) confirmed this idea. The ordering of how diverse the recommendation lists produced by the algorithms were, is in no way reflected in the general user satisfaction, usefulness, or trust of the system.

	Q1 accuracy	Q2 familiarity	Q4 novelty	Q5 diversity	Q7 transparency	Q8 satisfaction	Q10 trust	Q13 usefulness
Q1 accuracy	1	**.431**	**.459**	.012	**-.731**	**.767**	**.783**	**.718**
Q2 familiarity	**.431**	1	**.227**	.036	**-.405**	**.387**	**.429**	**.415**
Q4 novelty	**.459**	**.227**	1	**-.037**	**-.424**	**.496**	**.516**	**.542**
Q5 diversity	.012	.036	-.037	1	0.16	-.008	.001	-.096
Q7 transparency	**-.731**	**-.405**	**-.424**	.016	1	**-.722**	**-.707**	**-.622**
Q8 satisfaction	**.767**	**.387**	**.496**	-.008	**-.722**	1	**.829**	**.712**
Q10 trust	**.783**	**.429**	**.516**	.001	**-.707**	**.829**	1	**.725**
Q13 usefulness	**.718**	**.415**	**.542**	-.096	**-.622**	**.712**	**.725**	1

Table 3.6: The correlation matrix for the answers to the 8 most relevant questions on the online questionnaire of the user-centric evaluation.

To gain some deeper insight into the influence of the qualitative attributes towards each other, a simple linear regression analysis was performed. By trying to predict a qualitative attribute by using all the other ones as input to the regres-

sion function, a hint of causality may be revealed. Multiple stepwise regression analysis was used with bidirectional elimination: a combination of the forward selection approach, which step by step tries to add new variables (that have the highest marginal influence on the dependent variable) to its model, and the backward elimination approach, which step by step tries to remove the variables (with lowest marginal influence on the dependent variable) from the model. The following lines express the regression results. The qualitative attributes that were added to the model as predictive variables are indicated by means of an arrow notation and ordered (in descending order) according to their influence on the dependent variable. Between brackets we also noted the coefficient of determination, R^2. This coefficient indicates what percentage of the variance in the dependent variable can be explained by the model. R^2 will be 1 for a perfect fit and 0 if no linear relationship can be found.

accuracy $\leftarrow$ trust, transparency, usefulness, satisfaction $(R^2 = 0.7131)$

familiarity $\leftarrow$ trust, usefulness, transparency $(R^2 = 0.2195)$

novelty $\leftarrow$ usefulness, trust $(R^2 = 0.3260)$

diversity $\leftarrow$ usefulness, accuracy $(R^2 = 0.0230)$

transparency $\leftarrow$ accuracy, satisfaction, trust, familiarity $(R^2 = 0.6095)$

satisfaction $\leftarrow$ trust, accuracy, transparency, usefulness $(R^2 = 0.7470)$

trust $\leftarrow$ satisfaction, accuracy, usefulness, novelty, transparency, familiarity $(R^2 = 0.7625)$

usefulness $\leftarrow$ accuracy, trust, novelty, satisfaction, diversity, familiarity $(R^2 = 0.6395)$

The most interesting regression result is the line in which satisfaction (Q8) is predicted by the accuracy (Q1), transparency (Q7), trust (Q10), and usefulness (Q13). This result further strengthens our belief that accuracy (Q1) and transparency (Q7) are the main influencers of user satisfaction in our experiment (we consider trust (Q10) and usefulness (Q13) rather as results of respectively transparency and accuracy than real influencers of satisfaction, but they are of course also related). This regression model can also clarify the low performance of the SVD recommender regarding user satisfaction: the low transparency of the SVD recommender has a negative influence on the user satisfaction.

Because of the low and insignificant correlations between the diversity and the other qualitative attributes (Table 3.6), the regression model for the diversity (Q5) has a very low coefficient of determination $(R^2 = 0.0230)$. As a result, the variance in the diversity can only for a small fraction be explained in terms of the other qualitative attributes.

3.3.3 Offline evaluation

In addition to this online and user-centric experiment, an offline analysis was performed to compare the real, subjective opinions of the test subjects (originating from the online experiment) with the measured objective accuracy (of the offline analysis). For the offline analysis, recommendations were calculated on a training set containing 80% of the samples, which were randomly sampled from the collected feedback in the experiment. Using the leftover 20% as the test set, the accuracy of every algorithm was calculated over all users in terms of precision, recall, and F1-measure (Table 3.7). To average out any random effects, this procedure was repeated 10 times, each time with a different random partitioning of the data in training set and test set.

In this experiment, over 30,000 items were available for recommendation and on average only 22 items were consumed per user. Because of this extreme sparse consumption matrix, low values are obtained for the precision, recall, and F1-measure in the offline evaluation.

Algorithm	Precision (%)	Recall (%)	F1 (%)
CB	0.462	2.109	0.758
UBCF	1.359	4.817	2.119
Hybrid	1.173	4.377	1.850
SVD	0.573	2.272	0.915
Rand	0.003	0.015	0.005

Table 3.7: The accuracy of the recommendation algorithms in terms of precision, recall, and F1-measure based on an offline analysis

By comparing the offline and online results in our experiment, a small difference in the ranking of the algorithms can be noticed. In terms of precision, recall, and F1, the UBCF approach obtained the best results, followed by the Hybrid, SVD, CB, and Rand algorithm. Whereas the Hybrid approach performed best in the online analysis, this is not the case for the offline tests.

Note also that SVD and CB have swapped places in the offline ranking, compared to the ranking based on the question regarding accuracy of the online experiment. So according to the results of the offline analysis, SVD showed to be slightly better at predicting user behaviour than the CB algorithm. A possible explanation (for the inverse online results) is that test subjects in the online evaluation may have valued the transparency of the CB and Hybrid algorithm higher than its (objective) accuracy.

The results of our offline evaluation further underline the shortcomings of these offline procedures that only evaluate the prediction of user behaviour, thereby ne-

glecting the influence of the recommender system on the user behaviour and other qualitative attributes that contribute to the user experience.

It would have been interesting to be able to correlate the accuracy values obtained by the offline analysis with the subjective accuracy values obtained by the questionnaire on a user level. However, the offline evaluation showed very fluctuating results with on the one hand test subjects with close to zero precision and on the other hand some test subjects with relatively high precision values. As a result, the correlations between the results of the offline and online evaluation were not significant on a user level.

3.4 Conclusions

This chapter presented the results from a user-centric evaluation of recommendation algorithms, performed in the context of a Belgian cultural event website. The experiment evaluated the user experience of five commonly-used recommendation algorithms in terms of several qualitative attributes. Since offline evaluation metrics are inadequate for evaluating subjective characteristics such as usefulness and trust, an online, user-centric experiment was chosen as evaluation procedure.

In this experiment, both implicit and explicit feedback data were logged in the form of weighted user interactions with the event website over a period of 41 days. Given the availability of implicit and explicit feedback as well as item metadata, there were no restrictions regarding the algorithms that could be used. Implicit feedback that was logged before an event took place allowed CF algorithms to recommend events before the start date; and the availability of item metadata enabled CB approaches. Only in this ideal situation, a hybrid (CB+CF) algorithm can be used to generate recommendations in the context of an event information system.

Each of the test subjects in the experiment received a list of eight recommendations generated by one of the five algorithms. Subsequently, the test subjects were asked to fill in an online questionnaire that addressed the qualitative aspects of their recommendation list: accuracy, familiarity, novelty, diversity, transparency, satisfaction, trust, and usefulness.

Results clearly showed that the Hybrid algorithm, which combines the recommendations of the CB and UBCF algorithm, outperforms (or is equally as good in the case of question Q2 and the UBCF algorithm) every other algorithm except for the diversity aspect. In terms of diversity the random recommendations turned out best, which of course makes perfectly good sense. The runner-up for best algorithm in terms of qualitative aspects would definitely be the UBCF algorithm followed by the CB algorithm. This comes as no surprise considering that the Hybrid algorithm is mere a combination of these UBCF and CB algorithms. Since the UBCF algorithm is second best, it looks like this algorithm is the most responsible for the success of the Hybrid. While the weights of both algorithms were equal

in this experiment (i.e., the four best recommendations of each list were selected to be combined in the Hybrid list), it would be interesting to see how the results evolve if these weights would be tuned more in favour of the CF approach (e.g., $5\,UBCF + 3\,CB$).

Inspection of the correlation values between the answers of the questions revealed that diversity is in no way correlated with user satisfaction, trust, or any other qualitative aspect that was investigated. In contrast, the recommendation accuracy and transparency are qualitative aspects that are highly correlated with the user satisfaction and turned out to be influential predictors of the user satisfaction in the regression analysis.

The SVD algorithm came out last in the ranking of the algorithms and was statistically even indistinguishable from the random recommender for most of the questions except for again the diversity question (Q5). A histogram of the values for SVD and question Q8 (satisfaction) puts this into context by revealing an almost black and white opinion pattern expressed by the test subjects in the experiment. Moreover, since the user satisfaction is highly influenced by the transparency of the recommendations, the limited transparency of SVD might be another reason for the low subjective evaluations in the experiment.

References

[1] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. *Evaluating collaborative filtering recommender systems*. ACM Transactions on Information Systems, 22(1):5–53, 2004.

[2] S. M. McNee, J. Riedl, and J. A. Konstan. *Being accurate is not enough: how accuracy metrics have hurt recommender systems*. In CHI '06 extended abstracts on Human factors in computing systems, CHI EA '06, pages 1097–1101, New York, NY, USA, 2006. ACM.

[3] E. Campochiaro, R. Casatta, P. Cremonesi, and R. Turrin. *Do Metrics Make Recommender Algorithms?* In Proceedings of the 2009 International Conference on Advanced Information Networking and Applications Workshops, WAINA '09, pages 648–653, Washington, DC, USA, May 2009. IEEE Computer Society.

[4] C. Hayes, P. Massa, P. Avesani, and P. Cunningham. *An on-line evaluation framework for recommender systems*. In Workshop on Personalization and Recommendation in E-Commerce. Trinity College Dublin, Department of Computer Science, 2002.

[5] X. Guo, G. Zhang, E. Chew, and S. Burdon. *A hybrid recommendation approach for one-and-only items*. In Proceedings of the 18th Australian Joint

conference on Advances in Artificial Intelligence, AI'05, pages 457–466. Springer-Verlag Berlin Heidelberg, 2005.

[6] D. H. Lee. *PITTCULT: trust-based cultural event recommender*. In Proceedings of the 2008 ACM conference on Recommender systems, RecSys '08, pages 311–314, New York, NY, USA, 2008. ACM.

[7] R. Klamma, P. Cuong, and Y. Cao. *You Never Walk Alone: Recommending Academic Events Based on Social Network Analysis*. In J. Zhou, editor, Complex Sciences, volume 4 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 657–670. Springer Berlin Heidelberg, 2009.

[8] C. Cornelis, X. Guo, J. Lu, and G. Zhang. *A Fuzzy Relational Approach to Event Recommendation*. In Proceedings of the Indian International Conference on Artificial Intelligence: IICAI'05, pages 2231–2242, 2005.

[9] S. Dooms, T. De Pessemier, and L. Martens. *An Online Evaluation of Explicit Feedback mechanisms for Recommender Systems*. In Proceedings of the 7th International Conference on Web Information Systems and Technologies (WEBIST), 2011.

[10] D. Bollen, B. P. Knijnenburg, M. C. Willemsen, and M. Graus. *Understanding choice overload in recommender systems*. In Proceedings of the fourth ACM conference on Recommender systems, RecSys '10, pages 63–70, New York, NY, USA, 2010. ACM.

[11] Telematica Instituut / Novay. *Duine Framework*, 2009. Online available at `http://duineframework.org/`.

[12] J. S. Breese, D. Heckerman, and C. Kadie. *Empirical analysis of predictive algorithms for collaborative filtering*. In Proceedings of the fourteenth conference on Uncertainty in artificial intelligence, UAI'98, pages 43–52, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[13] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. T. Riedl. *Application of Dimensionality Reduction in Recommender System – A Case Study*. In ACM WebKDD Workshop, 2000.

[14] The Apache Software Foundation. *Apache Mahout*, 2012. Online available at `http://mahout.apache.org/`.

[15] P. Pu, L. Chen, and R. Hu. *A user-centric evaluation framework for recommender systems*. In Proceedings of the fifth ACM conference on Recommender systems, RecSys '11, pages 157–164, New York, NY, USA, 2011. ACM.

[16] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. McGraw-Hill, fifth edition, 2005.

[17] J. J. Higgins. *Introduction to Modern Nonparametric Statistics.* Thomson Brooks/Cole, first edition, 2004.

4

Group recommendations: considering multiple stakeholders

4.1 Introduction

In recent years recommender systems have become the common tool to handle the information overload problem of educational and informative web sites, content delivery systems, and online shops. Although most recommender systems make suggestions for individual users, in many circumstances the selected items (e.g., movies) are not intended for personal usage but rather for consumption in group. In these circumstances, the suggestions should be tailored to the entire group, which is composed of multiple stakeholders, to ensure maximum satisfaction of each individual user and the group as a whole. In this chapter, group recommendations are evaluated in terms of various qualitative aspects via an offline analysis. Moreover, a typical use case for group recommendations is discussed: a recommender system for audiovisual content that generates suggestions for groups of people (such as families or friends) in the home environment.

4.2 Test setup

4.2.1 Goals of the study

The main aim of this research is to find the optimal group recommendations approach for suggesting audiovisual content in the home environment. For this

scenario, we evaluate the group recommendations generated by various combinations of recommendation algorithms and aggregation strategies (i.e., strategies to convert traditional recommendation algorithms into group recommendation algorithms). The results show that the aggregation strategy which produces the most accurate results is depending on the algorithm that is used for generating individual recommendations. As in chapter 3, recommendations are not only assessed based on accuracy, but also on other qualitative aspects that are important for users such as diversity, coverage, and serendipity. Also these qualitative aspects of the group recommendations are to a large extent dependent on the used aggregation strategy and recommendation algorithm. Consequently for (commercial) group recommender systems, the aggregation strategy and algorithm have to be chosen carefully in order to optimise the desired qualitative aspects of the group recommendations.

In addition, this chapter investigates the influence of the size and composition of the group on the quality of the recommendations. In terms of group size, we test the hypothesis that the accuracy of the group recommendations decreases as the group size increases, since mediating the potentially contrasting preferences of the group members becomes more difficult for larger groups. In terms of group composition, intuition suggests that finding group recommendations that satisfy all group members should be easier as the group members are more similar to each other (i.e., have similar preferences). Therefore, we test the hypothesis that the accuracy of the group recommendations increases as the similarity between members of the group increases. Finally, this chapter proposes a combination of aggregation strategies which outperforms each individual strategy in terms of accuracy.

4.2.2 Group recommendations use case: a content delivery system for the home environment

This section provides an overview of the content delivery system for which we search the most effective group recommendation approach. Since the effectiveness of the recommendation algorithm can be dependent on the application domain and the system in which it is applied, we believe it is important to describe the features and functionally of this content delivery system.

The content delivery system provides a group of friends or a family suggestions for videos and songs originating from their joint collection of content items. So, the content delivery system has a key role in organizing, managing, delivering, and suggesting content that is available in the home network of the users. The system first aggregates the content of the group members from different sources in the home network (e.g., external hard drives, recorders, etc.) or even from sources in the home network of friends or relatives (if they gave permission to share their

content). Then, the system provides an overview of the users' joint collection of content items (songs and videos). If users select one of these content items, more information about the item is displayed (such as genres, actors, director, ...) together with a list of the most similar items.

Besides by browsing through the content items, users can find and explore content through their (personal) suggestions. These suggestions are calculated based on the preferences of the current group of users of the system. The content items and recommendations can be filtered based on genre to acquire a more specific selection of the content collection. E.g., users can query the system for recommendations of items in the genre 'Drama'. The quality of these group recommendations is further investigated in Section 4.3 of this paper.

Because of the limited hardware capabilities of the in-home device that runs the content delivery application, these recommendations are calculated by an external recommendation service and queried by the in-home device whenever needed. Since recommendations are calculated outside the home environment, in an external recommendation service which gathers the feedback of all users (of different homes), algorithms that take advantage of the community knowledge, such as Collaborative Filtering (CF), are applicable.

Subsequently, users can select a content item for playback on the desired device in the home environment (e.g., the television set). This interaction and viewing behaviour (play, pause, stop, ...) is logged as implicit feedback for the recommender system. Besides this implicit feedback, users can provide explicit feedback on individual items by the 'thumbs up' and 'thumbs down' icons or on genres, actors, and directors of the movie by selecting these attributes in the interface. Figure 4.1 illustrates this functionality of the content delivery system with a screenshot of the user interface.

Users can easily create or change their group according to the current situation in the home. E.g., a group can be composed for the family members that are planning to watch a movie. Besides adding users to a group or removing users from a group, a personal *importance weight* can be assigned to each member of the group. These weights can be used to specify the impact of each member's preferences on the group recommendations. This way, users can state for example that older people of the group (such as parents) have more influence on the recommendations than younger people (such as children). Three options are possible for these weights: a high, a low, and a neutral importance. The aggregation method of the group recommendation strategy takes these importance weights into account during the calculation of the group recommendation list (Section 4.3.1).

Changing the group composition or the importance weights has an immediate impact on the group recommendations which are shown in the interface. To enable these immediate adjustments to the recommendation list, recommendations

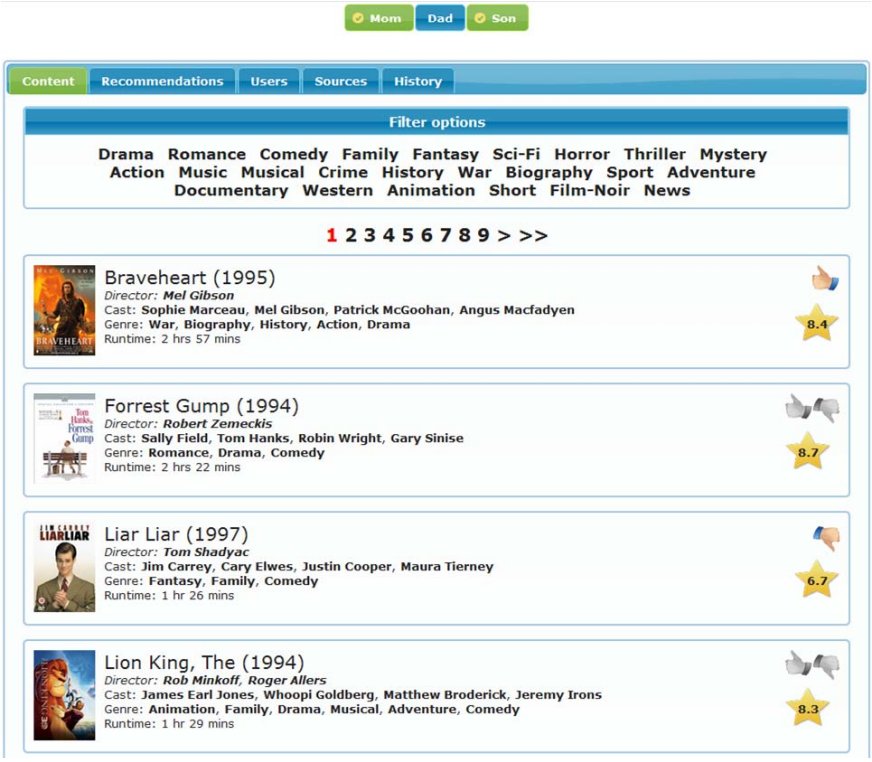


Figure 4.1: A screenshot of the content delivery system showing the current users composing a group (on top), the lists of content items, and the rating mechanism

are precalculated for every combination of group composition and importance weights. Given the small number of group members in a typical home environment (e.g., the family members) and the limited options for the importance weights (3 possible values), the total number of group combinations remains limited, so that the computation load is still acceptable.

4.2.3 Procedure

4.2.3.1 Evaluation method

To find the optimal group recommendation approach for the proposed system of Section 4.2.2, the effectiveness of the different aggregation strategies has to be measured for various state-of-the art recommendation algorithms and different compositions and sizes of the group. However, a major issue in the domain of group recommender systems is the evaluation of the effectiveness, i.e., comparing the generated recommendations for a group with the true preferences of the group.

Performing online evaluations or interviewing groups can be partial solutions but are not feasible on a large scale or to extensively test various combinations of alternative configurations. For example, in Section 4.3.2, five recommendation algorithms in combination with two aggregation strategies are evaluated for twelve different group sizes, thereby leading to 120 different setups of the experiment. In addition, Section 4.3.3 evaluates these five algorithms and two aggregation strategies for twenty additional group compositions with a varying similarity between the group members. This requires an additional number of 200 configurations. Furthermore, a data set with ratings originating from groups of people is, according to our knowledge, not available for research purposes. Therefore, we are forced to perform an offline evaluation, in which groups are sampled from the users of a traditional single-user data set, as was done by Baltrunas et al. [1].

In the literature, group recommendations have been evaluated several times by using a simulated data set with groups of users. Baltrunas et al. [1] used the MovieLens data set to simulate groups of different sizes (2, 3, 4, 8) and different degrees of similarity (high, random) with the aim of evaluating the effectiveness of group recommendations. Chen et al. [2] also used the MovieLens data set and simulated groups by randomly selecting the members of the group to evaluate their proposed group recommendation algorithm. They simulated group ratings by calculating a weighted average of the group members' ratings based on the users' opinion importance parameter. Quijano-Sánchez et al. [3] used synthetically generated data to simulate groups of people in order to test the accuracy of group recommendations for movies. In addition to this offline evaluation, they conducted an experiment with real users to validate the results obtained with the synthetic groups. To measure the accuracy of the group recommendations in the online experiment, they created groups of participants and asked them to pretend that they are going to the cinema together. One of the main conclusions of their study was that it is possible to realize trustworthy experiments with synthetic data, as the online user test confirmed the results of the experiment with synthetic data. This conclusion justifies the use of an offline evaluation with synthetic groups to evaluate the group recommendations in our experiment.

The offline evaluation of our experiment is based on the traditional procedure of dividing the data set in two parts: the training set, which is used as input for the algorithms to generate the recommendations, and the test set, which is used to evaluate the recommendations. In this experiment, we ordered the ratings chronologically and assigned the oldest 60% to the training set and the most recent 40% to the test set, as this reflects a realistic scenario the best. So, the ratings provided before a specific point in time are available as input for the recommender, whereas the ratings provided after that point in time are only used to evaluate the recommendations and not to train the recommender.

The evaluation procedure of the group recommendations, as proposed by Baltrunas et al. [1], is performed as follows. Firstly, artificial groups are composed by selecting random users from the data set. All users are assigned to one group of a predefined size. Secondly, group recommendations are generated for each of these groups based on the group members' ratings in the training set. Since group recommendations are intended to be consumed in group and to suit simultaneously the preferences of all members of the group, all members receive the same recommendation list. Thirdly, the recommendations are evaluated individually as in the classical single-user case, by comparing (the rankings of) the recommendations with (the rankings of) the items in the test set of the user.

4.2.3.2 Data set

The group recommendations generated by using various combinations of alternative algorithms, aggregation strategies, and group sizes are evaluated offline using the MovieLens (100K) data set [4]. This data set contains information about 1682 popular movies, including 100000 evaluations on a 5-point rating scale of 943 users.

Using this data set in the calculation process of the recommender service can also help to overcome the cold start problem for the first users of our system (as presented in Section 4.2.2). Therefore, the explicit and implicit feedback provided by the (future) users of our system will be converted to the 5-point rating scale of the MovieLens system. This way, the combined data set (MovieLens data + feedback from our users) enables CF algorithms to find neighbours for the new users of our system and generate accurate recommendations based on the community knowledge of the MovieLens data set.

Before calculating the recommendations, the data set is first transformed to optimally estimate the preferences of the users. The user's ratings are normalized by subtracting the user's mean rating (i.e., μ) and dividing this difference by the standard deviation of the user's ratings (i.e., σ).

$$r_{norm} = \frac{r - \mu}{\sigma} \tag{4.1}$$

This normalization is required to compensate for very enthusiastic users giving only positive ratings or very critical users who mainly provide negative feedback. Some similarity metrics, such as the Pearson correlation, consider the fact that users are different with respect to how they interpret the rating scale, thereby making the normalization process unnecessary for calculating similarities. However, normalizing the ratings is still necessary if the ratings of the group members are aggregated into a group rating before the similarities are calculated [5].

4.2.3.3 Algorithms

The main goal of this research is not to develop new recommendation algorithms, but rather to investigate how effective group recommendations can be generated by aggregating the group members' data and using existing recommendation algorithms. Therefore, different group recommendation strategies are investigated by using a number of state-of-the-art recommendation algorithms. As in chapter 3, a Content-Based (CB) recommendation algorithm, a nearest neighbour CF technique, a hybrid CF-CB algorithm (Hybrid), and a recommendation algorithm based on Singular Value Decomposition (SVD) are evaluated. Details about these algorithms are provided in Section 3.2.2.2.

Since the offline evaluation procedure enables the extensive testing of various algorithms, the User-Based Collaborative Filtering (UBCF) as well as the Item-Based Collaborative Filtering (IBCF) approach are used to produce group recommendations in this evaluation. Experimental evaluations showed that these IBCF algorithms are faster than the traditional user-neighbourhood based recommender systems and provide recommendations with comparable or better quality if sufficient data is available [6]. Because of this, the IBCF algorithm is chosen (instead of the UBCF algorithm) as the complementary recommender of the CB algorithm in the Hybrid recommender. The user-centric evaluation (of chapter 3) comparing the different algorithms based on various characteristics (including accuracy, novelty, diversity, satisfaction, and trust) showed that the Hybrid combination of CF and CB recommendations outperforms both individual algorithms on almost every qualitative metric. The SVD recommender is configured to use 19 features, i.e., the number of genres in the MovieLens data set, and the number of iterations is set at 50.

To compare the results of the different recommenders, the *most-popular recommender* was introduced as a baseline algorithm. This recommender generates for every user or group always the same static list of the most-popular items in the system, regardless the ratings or activity of the user or group. The popularity of an item is estimated by the number of ratings and the average of the ratings the item received (in the training set). Using the community knowledge available in the MovieLens data set, it was possible to use this most-popular recommender as an improved baseline algorithm compared to the random recommender of Section 3.2.2.2.

4.2.4 Evaluation metrics

4.2.4.1 Accuracy

Since predicting the effective rating value is less important for the use case of group recommendations for audiovisual content (Section 4.2.2), but rather the correct ordering of content items according to the group's preferences, a ranking met-

ric is used to assess the accuracy of the recommendations. So, the group recommendations are evaluated based on the individual ratings in the test set using the normalized Discounted Cumulative Gain (nDCG), as explained in Section 1.2.1.3. We adopted the suggestion of Baltrunas et al. [1] to compute the nDCG based on the projection of the recommendation list on the test set of the user. In this experiment, we opted for $n = 5$ as the number of recommendations, since this is a realistic length for a manageable recommendation list in a TV interface (Section 4.2.2). After calculating the nDCG for each individual user, the mean nDCG over all users of the data set is calculated as an overall measure of efficiency. The resulting nDCG ranges between 0 and 1; higher values indicate more accurate group recommendations.

This accuracy evaluation, which is based on combining individual users into synthetic groups, has a limitation compared to an evaluation with real groups of users. There is no way of finding out how satisfied individuals really would be with the group recommendations (in the way a real group could be asked, and real group members would take the feelings of others in the group into account). So for the offline evaluation of group recommendations based on a data set with ratings of individuals, the only possible resort is to approximate the preferences of the user being in a group, by the preferences of the user evaluating the content individually. Despite this limitation, evaluating the accuracy of group recommendations by generating synthetic groups has already proven its usefulness in previous research [1–3]. For the other quality metrics, such as diversity, coverage, and serendipity, the evaluation methodology based on synthetic groups is not a limitation.

4.2.4.2 Diversity

For the use case of our recommender system for audiovisual content (Section 4.2.2), it is desirable that the content items of the recommendation list are covering different genres. Therefore, we measure the item-item similarity between recommended items based on the genres describing these content items. So, the item-item similarity of two content items c_i and c_j is measured by comparing the set of genres describing the first item $c_{i_{genres}}$, to the set of genres describing the second item $c_{j_{genres}}$, using the Jaccard similarity coefficient. Besides the genres, also actors and directors could be included as keywords in the sets that describe an item. However because of the limited overlap of actors and directors compared to the overlap of genres for pairs of content items, the inclusion of actors and directors had no significant effect on the diversity for this experiment.

The Jaccard similarity coefficient calculates the similarity of two sets by the ratio of the intersection of the sets and the union of the sets:

$$Sim(c_i, c_j) = \frac{c_{i_{genres}} \cap c_{j_{genres}}}{c_{i_{genres}} \cup c_{j_{genres}}} \tag{4.2}$$

Subsequently, the *intra-list similarity*, i.e., a measure for the similarity of all items within a recommendation list [7], is estimated by the arithmetic mean of the item-item similarity of every couple of items in the list.

$$IntraList\ Similarity = \frac{2 \cdot \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Sim(c_i, c_j)}{n \cdot (n-1)} \tag{4.3}$$

This intra-list similarity is calculated for the recommendations of every user and the mean over all users is calculated to obtain a global value for the similarity of items within a recommendation list. Finally the diversity of the recommended items is calculated by subtracting this mean intra-list similarity from 1.

$$ListDiversity = 1 - mean(IntraList\ Similarity) \tag{4.4}$$

Because of the definition of the Jaccard similarity coefficient, the mean intra-list similarity ranges between 0 and 1. So, the diversity of the recommendation list varies from 0 (very similar recommendations) to 1 (very diverse recommendations).

4.2.4.3 Coverage

As suggested by Herlocker et al. [8], the catalog coverage is measured by taking the union of the top-N recommendations for each user in the population. In case the users are partitioned into groups, and group recommendations are calculated instead of individual recommendations, we measure the catalog coverage based on the union of the top-N recommendations for each of these groups. Subsequently, the cardinality of this set (i.e., the number of items in this union) is divided by the number of items in the catalog of the system to obtain the catalog coverage.

Let us denote $rec(u_i)$ as the recommendation list of user u_i. The number of users for which recommendations are generated is k. Let cat be the set of all available items in the system, and $|cat|$ is the cardinality of this set. Then the catalog coverage can be measured as follows [9]:

$$CatalogCoverage = \frac{|\cup_{i=1\ldots k}\ rec(u_i)|}{|cat|} \tag{4.5}$$

The values of the catalog coverage range from 0, meaning that the recommender suggests none of the items, to 1, meaning that all items of the catalog are recommended to at least one user. Catalog coverage is usually measured on a specific set of recommendations, at a single point in time [8]. For instance in this research, it is measured based on the union of the top-5 recommendations, calculated based on the training set, for each user or group in the population.

4.2.4.4 Serendipity

Shani and Gunawardana [10] proposed to estimate the serendipity by a distance measurement between a recommended content item c_i, and the set of content items in the profile of the user u, i.e., the items that the user has previously watched, bought, or consumed. Although this metric is explained in the context of a book recommender and considers the authors of the books, it can easily be generalized to estimate the serendipity of any type of content item based on the metadata attributes of that item (e.g., the genres). So for the evaluation of the group recommendations, we used the following generalization of the metric of Shani and Gunawardana to estimate the serendipity of the recommended movies based on their genres (Section 4.3.2.5 and 4.3.3.5).

Let us denote $g(c_i)$ as the genre or set of genres categorizing the content item c_i. Let $N_u(g)$ be the number of items in the profile of the user u, that are described by the genre g. If g is a set of genres consisting of $\{g_1, g_2, \ldots, g_l\}$, than $N_u(g)$ is the mean of all $N_u(g_i)$ calculated over all genres in the set, $i = 1, \ldots l$. The number of items in the user's profile that are categorized by the user's most chosen genre is represented by $N_u(g_{max})$:

$$N_u(g_{max}) = \max_i(N_u(g_i)) \tag{4.6}$$

The relevance of a content item c_i can be denoted by the boolean function $isRelevant(c_i) \in \{0, 1\}$, where $isRelevant(c_i) = 1$ means that c_i is interesting for the user, and $isRelevant(c_i) = 0$ means that it is not [11]. We consider all items in the test set that received a rating of 3, 4, or 5 stars (on a 5-point scale star-rating mechanism) from the user as relevant for that user. In contrast, items in the test set that received a poor rating (1 or 2 stars) are considered as uninteresting or unrelevant for the user. The personal relevance of an item that is not yet rated by that person is unknown and difficult to judge. Therefore, we give these unrated items the benefit of the doubt and consider them as potentially relevant for the user, $isRelevant(c_i) = 1$. This favours algorithms which generate recommendations for new, unknown, or niche items, in contrast to the popular, commonly rated items. Finally, the serendipity of a recommended content item c_i can be calculated as follows:

$$Serendipity(c_i) = \frac{1 + N_u(g_{max}) - N_u(g_i)}{1 + N_u(g_{max})} \cdot isRelevant(c_i) \tag{4.7}$$

The values of the serendipity range from 0, meaning that the recommender only suggests obvious or unrelevant items, to 1, meaning that all the recommended items are relevant and surprising. Next, the *list-serendipity* is estimated by the mean of the serendipity of every item in the recommendation list. The mean of the list-serendipity of each user's recommendation list is used as a global measure for the serendipity of a recommendation algorithm in Section 4.3.2.5 and 4.3.3.5.

4.3 Results

In this section, we discuss the results of the evaluation of the group recommendations calculated by different algorithms. First, the influence of the data aggregation method on the accuracy of the group recommendations is discussed in Subsection 4.3.1. Subsequently, this study evaluates the recommendations for groups of a varying size (Section 4.3.2) and a varying composition, i.e., randomly composed groups versus groups with like-minded members (Section 4.3.3). This evaluation is based on various quality metrics (accuracy, diversity, coverage, and serendipity) as discussed in Section 4.2.4, in order to assess the recommendations on different aspects. Finally, Section 4.3.3 discusses how aggregation strategies can be combined in order to obtain more accurate group recommendations.

4.3.1 Influence of the data aggregation method

4.3.1.1 Data aggregation methods

As explained in Section 1.3.1, the (data) aggregation method is the mathematical function that determines how the individual recommendation lists of group member are combined into group recommendations in case of the aggregating recommendations strategy, or how the individual group members' preferences are combined into a group preference in case of the aggregating preferences strategy.

So, in case of the aggregating recommendations strategy, a standard recommendation algorithm is used to calculate a prediction of the user's rating for each content item in the system and for each user of the group. Next, the content items can be sorted by this prediction value in a descending order to obtain a list of recommendations for each individual user. To obtain group recommendations, the individual recommendations of the group members are aggregated by combining the prediction values of each group member's recommendation list according to the data aggregation method. Subsequently, the recommended items are sorted by this aggregated prediction value in descending order. Finally, the group recommendation list is obtained by keeping the top-N items.

In case of the aggregating preferences strategy, the members' individual preferences are aggregated into a group preference by combining the members' rating for each item according to the data aggregation method and using this aggregated result as a group rating. Subsequently, group recommendations are calculated based on these group ratings using a standard recommendation algorithm. Again, only the top-N recommendations are offered to the group.

A determining factor in the selection process of the aggregation method is the resulting accuracy of the group recommendations. Therefore, the influence of the aggregation method on the accuracy of the group recommendations is investigated

by comparing the following five aggregation methods, which have been proposed in literature [5].

4.3.1.1.1 Average (Avg) In case of the aggregating recommendations strategy, the first aggregation method, i.e., *average*, aggregates the individual recommendation lists by calculating the average of the prediction values of the members' ratings and uses this average as the prediction value for the group. In case of the aggregating preferences strategy, the average method aggregates the individual preferences by calculating the average of the members' ratings and uses this average as the group rating. Because this method aggregates preferences and recommendations in a desirable and intuitive way (as discussed in Section 4.3.1.4), and because this method corresponds to one of the ways in which a group of people naturally make choices [5], we used this aggregation method for the experiments of Section 4.3.2, 4.3.3, and 4.3.4.

If group members have an unequal importance weight, which reflects the situation that some users have more influence on the group recommendations than other users (Section 4.2.2), a weighted average can be used as aggregation method to take the relative importance of each group member into account. Unfortunately, the influence of the importance weights on the accuracy of the group recommendations could not be evaluated in the experiment of Section 4.3.1.2, since the data set that was used for this research does not contain these weights.

4.3.1.1.2 Average Without Misery (AvgWM) The idea of the *average without misery* method is to find the optimal decision for the group, without making some group members really unhappy with this decision. If the recommendations are aggregated, the average of the prediction values of each recommendation list is calculated. Items that have a prediction value below a certain threshold (in one of the recommendation lists) get a penalty or are excluded from the group recommendations. Then the recommended items are sorted in descending order based on this new prediction value. In our implementation, the threshold is set at 2, so if an item appears in the recommendation list of a member with a prediction value below 2, the prediction value in the recommendation list of the group is set to 1. This corresponds to disfavouring the item with respect to all other available items, thereby making it very unlikely to appear in the group recommendation list.

If the preferences are aggregated, the group rating for an item is the average of the ratings of the members for that item. However, items that are rated below a certain threshold by one of the members get a penalty. Also for this strategy, the threshold is set at 2; and the penalty rule converts an individual rating below this threshold into the group rating. So if at least one group member gives a rating of 1 star to an item (i.e., below the threshold of 2 stars), the group rating is 1, otherwise the group rating is the average of the members' ratings.

4.3.1.1.3 One user choice (One) The aggregation method that is referred to as *one user choice*, sometimes also called "most respected person" or "dictatorship", adopts the preferences of one user in the group. The idea is that one group member might be the user that makes the decision about what the group is going to choose without consulting the other group members. In our implementation, this user is chosen randomly from the group members. So in case of the aggregating recommendations strategy, the group's prediction value for an item is equal to the prediction value of a randomly-chosen member for that item. In case of the aggregating preferences strategy, the group's rating for an item is the rating of a randomly-chosen member for that item.

4.3.1.1.4 Least Misery (LM) The *least misery* aggregation method tries to minimize the "misery" for the group members. The idea is that the group is as happy as its least happy member. Therefore, the goal is to obtain at least a predefined level of satisfaction for all group members. This method is implemented as follows: if the recommendations are aggregated, the group's prediction value for an item is equal to the minimum of the prediction values of all group members for that item. If preferences are aggregated, the group's rating for an item is the minimum of the members' ratings for that item.

4.3.1.1.5 Most Pleasure (MP) The aggregation method called *most pleasure* tries to maximize the "pleasure" for (one of) the group members. This method tries to recommend alternately the items that one group member really likes, thereby not considering the preferences of other members. In case of the aggregating recommendations strategy, the group's prediction value for an item is equal to the maximum of the prediction values of all group members for that item. In case of the aggregating preferences strategy, the group's rating for an item is the maximum of the members' ratings for that item.

4.3.1.2 Aggregation method experiment

To investigate the influence of the data aggregation method on the accuracy of the group recommendations, group recommendations generated using each of these aggregation methods are compared via a series of experiments (Section 4.3.1.3). In these experiments, the groups are composed by selecting random users, meaning that no additional restrictions are imposed on the group or on the group members. To investigate the influence of the aggregation method separately from other parameters, the group size is fixed (at 2 or 5) in these experiments. For each algorithm, the two strategies to generate group recommendations (aggregating recommendations and aggregating preferences) are evaluated.

Since users are randomly combined into groups and the quality of group recommendations is depending on the composition of the groups, the quality metrics

slightly vary for each partitioning of the users into groups. (Except for the partitioning of the users into groups of 1 member, which is only possible in 1 way.) Therefore, the process of composing groups by taking a random selection of users is repeated 30 times and just as much measurements of the quality metric are performed. The mean of these 30 measurements is used as an estimation of the quality of the group recommendations and is visualized in the corresponding graph (Figures 4.2 and 4.3) (on the vertical axis) together with the 95% confidence intervals of the mean values. The used aggregation method is indicated on the horizontal axis. If two bars have non-overlapping confidence intervals, they are necessarily significantly different (but if they have overlapping confidence intervals, it is not necessarily true that they are not significantly different).

The bar series with the prefix "Rec" evaluate recommendation algorithms in combination with the aggregating recommendations strategy whereas the prefix "Pref" refers to the aggregating preferences strategy. For example, the bar series "PrefUBCF" stands for the group recommendations which are generated by combining the members' individual preferences using the aggregating preferences strategy and calculating recommendations for this aggregated profile using the user-based collaborative filtering algorithm.

The vertical axes of the graphs (Figures 4.2 and 4.3) cross the horizontal axes at the quality level of the most-popular recommender (i.e., $nDCG = 0.8722$), which is constant for the different group sizes and aggregation methods. This way, the bar charts show the relative improvement (or deterioration) of each algorithm with respect to the baseline quality of the most-popular recommender.

4.3.1.3 Accuracy influenced by the aggregation method

Figures 4.2 and 4.3 show the mean nDCG (calculated over all users) together with the 95% confidence interval of the mean nDCG, in relation to the recommendation algorithm, the aggregation strategy (aggregating preferences or aggregating recommendations), and the aggregation method. Figure 4.2 shows the accuracy of the group recommendations for groups of 2 members; whereas Figure 4.3 shows the accuracy for groups of 5 members.

As visible in Figure 4.2, the influence of the aggregation method on the accuracy of the group recommendations is largely dependent on the algorithm and aggregation strategy. E.g., the accuracy of the recommendations generated by the Hybrid recommender in combination with the aggregating preferences strategy (PrefHybrid), remains approximately constant over the different aggregation methods. In contrast, the accuracy of the recommendations generated by RecCB, significantly varies if different aggregation methods are used.

The aggregation method that produces the most accurate group recommendations depends on the used algorithm and aggregation strategy. E.g., the PrefCB combination produces the most accurate group recommendations if the MP method

is used. If the RecCB combination is used, the most accurate group recommendations are obtained by choosing LM as aggregation method. The PrefUBCF combination provides the best results together with the AvgWM method; and the RecHybrid combination generates the most accurate recommendations if the Avg method is used. Although the confidence intervals indicate that not all differences are significant, the results show that the choice of the best aggregation method is directly linked to the aggregation strategy and recommendation algorithm.

Figures 4.2 and 4.3 show that the Avg and AvgWM method generally provide the most accurate and also the most stable results. As expected, the 'one user choice (One)' method has poor results in combination with the aggregating recommendations strategy (Rec), especially with RecCB and RecUBCF. The selection of a prediction value from one random member for all recommended items has a drastic influence on the resulting group recommendations. On the other hand, selecting the ratings from a random member as group rating has less influence on the final recommendations, since this happens much earlier in the recommendation process.

The LM method leads to a decreased accuracy in combination with RecUBCF and the MP method generates less accurate recommendations if RecCB or RecUBCF is used. Again, the aggregation of recommendations, which happens late in the recommendations process, can have a serious impact on the accuracy of the group recommendations because the aggregation method does not sufficiently takes into account the preferences of all members.

Comparing Figures 4.2 and 4.3 confirms that the results for groups of 2 members are in line with the results for larger groups (e.g., 5 members per group): the optimal aggregation method has to be chosen based on the used recommendation algorithm and aggregation strategy. Moreover, the results of Figure 4.3 indicate that a sub-optimal aggregation method can have a dramatic impact on the accuracy of the recommendations, especially for larger group sizes. E.g., the accuracy of the recommendations obtained by using the aggregating recommendations strategy (Rec) and the one user choice (One) aggregation method, is significantly lower than the level of the horizontal axis, which indicates the accuracy of the list of the most popular items.

Although several other aggregation methods have been proposed in literature [5], the results of this experiment already indicate that 'one best' aggregation method, that generates the most accurate group recommendations for all combinations of aggregation strategy and algorithm, may not exist. So for an optimal group recommender system, the aggregation method has always to be chosen in combination with the recommendation algorithm and the aggregation strategy.

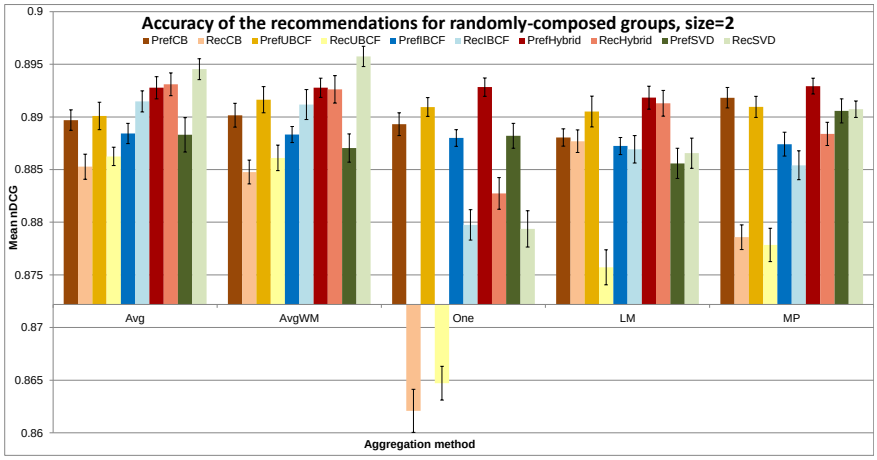


Figure 4.2: The accuracy of the group recommendations for groups of size = 2, generated by using different aggregation methods

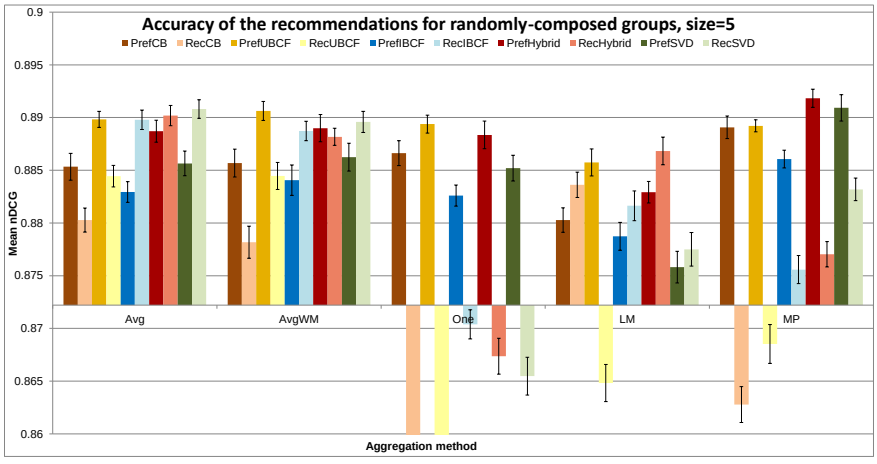


Figure 4.3: The accuracy of the group recommendations for groups of size = 5, generated by using different aggregation methods

4.3.1.4 Aggregation method selection

The context and application area in which group recommendations are required may also have an influence on the choice of the recommendation strategy and aggregation method. For example in a family context, meals or holiday destinations that are really disliked by one member of the family will often not be chosen for the group, regardless the opinion of the other family members. Different reasons for a strong aversion to a particular item may exist: a family member might be al-

lergic to a specific ingredient of the meal or a family member might be (physically) unable to travel to a specific holiday destination. During these joint decisions, a solidarity between the family members exists. So, a decision that leaves one or more family members very dissatisfied is likely to be considered undesirable, even if the average satisfaction is high [12]. Since these items are undesirable as a group recommendation, a *minimizing misery* approach such as the *average without misery* or *least misery* aggregation method [5] is appropriate in this context.

In the context of movies or music on the other hand, users might be more willing to watch or listen to something they dislike, if the other members of the group enjoy it. E.g., people may join their friends for watching a movie or listening to music because of the company, even if they do not like some of the movies or songs during the event. Users might be willing to renounce their personal preferences in order to *maximize the average satisfaction* of the group. As a result, the *average* function is a proper candidate as aggregation method. Moreover, research has shown this method to be one of the ways in which a group of people intuitively come to a group decision [5].

In this research, the different group recommendation strategies and algorithms were evaluated in the context of a recommender system for movies (and songs) in the home environment (Section 4.2.2). Because of the targeted application domain of the recommender (i.e., audiovisual content), the *average* function was chosen in Section 4.3.2, 4.3.3.1, and 4.3.4.1 to combine the individual recommendation lists in the case of the aggregating recommendations strategy and to combine the members' preferences in the case of the aggregating preferences strategy. By using the same aggregation method (i.e., average) for both aggregating the individual recommendation lists and aggregating the individual preferences, the accuracy of all strategies can be compared.

Moreover, the higher average performance of the Avg method compared to the AvgWM method (Section 4.3.1.3) was an additional argument to choose for the Avg aggregation method for our recommender system. E.g, the recommendations for groups of 5 members generated by RecCB are significantly better in combination with the Avg method than with the AvgWM method (statistical T-test: $t(58) = 2.17, p = 0.03 < 0.05$). Consequently, all experiments of Section 4.3.2, 4.3.3, and 4.3.4.1 rely on the *average* function to aggregate preferences or recommendations.

4.3.2 Influence of the group size

4.3.2.1 Group size experiment

The second series of experiments (Section 4.3.2.2, 4.3.2.3, 4.3.2.4, and 4.3.2.5) investigates the influence of the group size on the quality of the group recommendations. The group size is varying from 1 person per group (i.e., individual

recommendations) to 10 persons per group. Besides, the results are provided for very large group compositions (group sizes of 15 and 20 persons). In contrast to the first experiments, all the combinations of aggregation strategy and recommendation algorithm use the average (Avg) as aggregation method.

Just like in the first series of experiments, the groups are composed by selecting random users from the data set and the process of composing groups is repeated 30 times. So, each quality metric is calculated 30 times and the mean of these measurements is used as an estimation of the quality of the group recommendations. The graphs in Figures 4.4, 4.5, 4.6, and 4.7 show these mean values (on the vertical axis), as well as the 95% confidence intervals of the mean values; the group size is indicated on the horizontal axis. Again, the vertical axis of each figure crosses the horizontal axis at the quality level of the most-popular recommender and the prefix of the bar series denotes if the algorithm uses the aggregating recommendations strategy ("Rec") or the aggregating preferences strategy ("Pref").

4.3.2.2 Accuracy influenced by the group size

Figure 4.4 shows the mean nDCG (calculated over all users) together with the 95% confidence interval of the mean nDCG, in relation to the recommendation algorithm, aggregation strategy, and the group size. All bar series are significantly higher than the horizontal axis indicating the accuracy level of the most-popular recommender (i.e., $nDCG = 0.8722$). So each combination of algorithm, aggregation strategy, and group size shows an accuracy improvement with respect to the static list of most popular items, which proves the usefulness of group recommendations, even for large groups.

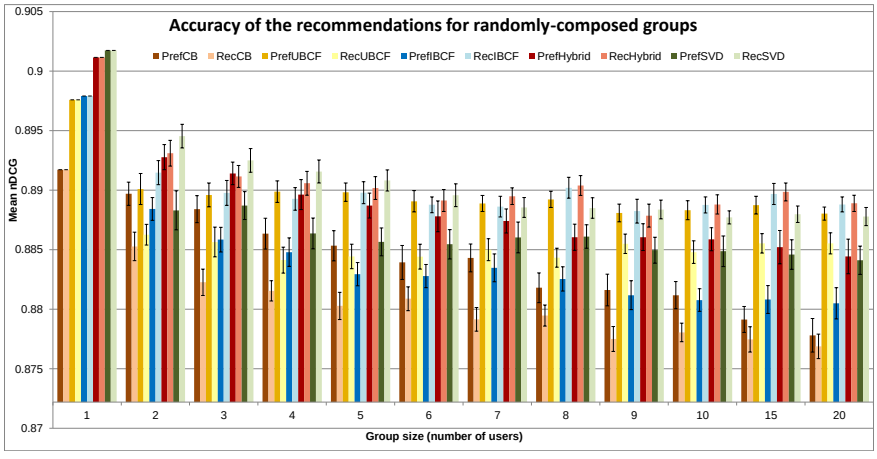


Figure 4.4: The accuracy of the group recommendations for randomly-composed groups of a varying group size

A comparison of the different algorithms of Figure 4.4 indicates that the SVD and Hybrid recommender produce the most accurate group recommendations for various group sizes. However, the difference in accuracy with UBCF and IBCF is small. In contrast, the CB recommender generates the least accurate group recommendations, which are nevertheless still significantly better than the list of most popular items.

As expected, Figure 4.4 shows for all algorithms a decreasing performance regarding the accuracy of the group recommendations as the group size increases. However, this decrease is not equally large for all algorithms: a large decrease is witnessed for PrefCB, RecCB, PrefIBCF, PrefHybrid, and RecSVD, whereas PrefUBCF, RecUBCF, RecIBCF, RecHybrid, and PrefSVD suffer only from a slight decrease in accuracy as the group size increases. A larger group signifies more members and more individual preferences to take into account during the recommendation process. Since the groups are randomly composed, members can have different or even opposite preferences. So for these random groups, recommending items that are interesting for all members becomes more difficult when the group size increases.

The comparison between the strategy that aggregates recommendations and the strategy that aggregates preferences provides another interesting finding. The aggregation strategy that provides the most accurate recommendations depends on the used algorithm. The CB and UBCF algorithm generate the most accurate group recommendations if the group members' preferences are aggregated, whereas the results of SVD and IBCF are most accurate if the members' recommendations are aggregated. The Hybrid recommender generates the most accurate recommendations in combination with the aggregating recommendations strategy, but the differences are not significant for small groups. Table 4.1 shows the results of the statistical T-tests comparing the mean accuracy of the recommendations generated by the two aggregation strategies for groups of five members. (Similar results are obtained for other group sizes.) The null hypothesis, H_0 = the mean accuracy of the recommendations generated by the aggregating preferences strategy is equal to the mean accuracy of the recommendations generated by the aggregating recommendations strategy. T-tests with a p-value below 0.05 indicate a significant difference between the two aggregation strategies, and are presented in bold.

A possible explanation for these differences in accuracy lies in the way in which the algorithm processes the data. The CB and UBCF algorithm create a user profile modelling the user's preferences in order to find items matching this profile (in the case of the CB algorithm) or to find users with similar preferences (in the case of UBCF). So for these algorithms, aggregating the members' preferences corresponds to aggregating the profile models of the group members. In contrast, the matrix decomposition of SVD and the item-item similarities of IBCF

Algorithm	t(58)	p-value
CB	5.03	**0.00**
UBCF	7.17	**0.00**
IBCF	-8.70	**0.00**
Hybrid	-1.77	0.08
SVD	-5.99	**0.00**

Table 4.1: Statistical T-test comparing the mean accuracy obtained by the two aggregation strategies for groups with size = 5

provide less insight into the preferences of the users or the aggregation of these preferences. The Hybrid recommender, which combines the IBCF and CB recommender, reflects the accuracy differences for the aggregation strategies of the underlying algorithms.

So, aggregating the preferences of the group members provides optimal results if the algorithm internally composes some kind of user profile holding the users' preferences, whereas aggregating the recommendations of the group members is a better option if the users' preferences are less transparent in the data structure of the algorithm. The internal modelling of the user profile can also explain why some combinations of algorithm and strategy (such as PrefSVD) deteriorate faster than others (such as PrefUBCF) as the group size increases. Consequently, if an existing recommender system for individuals is extended to a recommender system for groups, the aggregation strategy has to be chosen based on the utilized recommendation algorithm in order to maximize the efficiency of the group recommendations.

4.3.2.3 Diversity influenced by the group size

Figure 4.5 shows the mean list diversity (calculated over all users) together with the 95% confidence interval of the mean list diversity, in relation to the recommendation algorithm, aggregation strategy, and the group size.

The list diversity of the most-popular recommender is 0.72, which is indicated in Figure 4.5 by the level of the horizontal axis. Since the most-popular recommender is based on the consumption behaviour of the whole community, the suggestions consist of a set of dissimilar items covering different genres. As a result, the recommendation list generated by the most-popular recommender is rather diverse in comparison with the other algorithms such as CB and SVD.

The results reveal a clear ranking of the algorithms based on the list diversity. The CB recommender scores much worse than the most-popular recommender and produces the least diverse recommendation lists. This poor diversity is due to

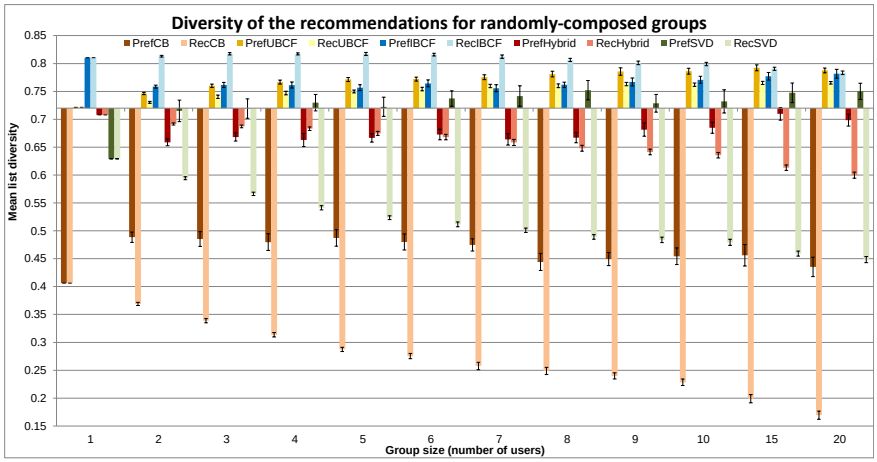


Figure 4.5: The diversity of the group recommendations for randomly-composed groups of a varying group size

the reasoning process of the CB recommender. E.g., if a user gave only positive evaluations to action movies in the past, the CB recommender will only suggest more action movies to this user. In this case, the recommendation list consists of all very similar items and as a result, it has a low list diversity. This is the well-known problem of 'over-specialization' of CB recommenders. One of the purposes of hybrid systems (comparing to CB systems) is to try to overcome this problem of over-specialization. Nevertheless because of the high similarity of the CB recommendations, also the Hybrid recommender provides a recommendation list that is less diverse than the most popular list.

The recommendations based on SVD are in most cases less diverse than the most popular items. Only the recommendations based on SVD which are generated for large groups by aggregating the members' preferences are more diverse than the most popular items. The low diversity of these recommendations might be due to the 'feature identification' of the SVD algorithm. The matrix decomposition of the algorithm reduces the user-item matrix into a smaller-dimensional space where highly correlated items (for example, movies of the same genre, same actor, ...) are captured as a single feature. Then, the resulting recommendations are characterized by the same features as the items that the user appreciated in the past.

So the CB recommender and to a lesser extent SVD can trap (individual) users in a 'similarity loop', only giving exceptionally similar recommendations of the same genre over and over again, without suggesting new or surprising items to the user. If the profile of an individual user is aggregated with the profile of another user, the resulting group profile can contain a greater variety of consumed items.

This is visual in the results of PrefCB and PrefSVD which show an increased list diversity when the group size grows from 1 individual user to a group of 2 members.

The algorithms based on CF generate more diverse recommendations than the most-popular recommender. The Pearson correlation metric for discovering similar users in the user-based approach (UBCF) or similar items in the item-based approach (IBCF) introduces the necessary diversity. E.g., the UBCF recommender can suggest a horror movie to a user who never rated a horror movie, because a similar user liked that horror movie. The most diverse recommendation list is obtained by using the IBCF recommender in combination with the aggregating recommendations strategy. So, the item matching process of IBCF using the Pearson correlation metric results in a very diverse set of recommendations.

For most algorithms and strategies, the diversity remains constant as the group size increases. Except for RecCB, RecSVD, and RecHybrid, the diversity decreases as the group size increases. The recommendation lists for individual users (group size = 1) generated by these algorithms consist of very similar items, and combining these recommendation lists stimulates this similarity.

When we compare the two aggregation strategies, SVD, UBCF and the CB recommender produce the most diverse recommendations if the preferences are aggregated whereas the group recommendations of IBCF are more diverse if the members' individual recommendation lists are aggregated. The Hybrid recommender follows the behaviour of the underlying algorithms and generates more diverse recommendations for small groups if recommendations are aggregated and for large groups if preferences are aggregated. Table 4.2 shows the results of the statistical T-tests comparing the mean diversity of the recommendations generated by the two aggregation strategies for groups of five members. (Similar results are obtained for other group sizes.) H_0 = the mean diversity of the recommendations generated by the aggregating preferences strategy is equal to the mean diversity of the recommendations generated by the aggregating recommendations strategy. T-tests with a p-value below 0.05 indicate a significant difference between the two aggregation strategies, and are presented in bold.

Algorithm	t(58)	p-value
CB	22.25	**0.00**
UBCF	8.06	**0.00**
IBCF	-17.48	**0.00**
Hybrid	-1.61	0.11
SVD	19.12	**0.00**

Table 4.2: Statistical T-test comparing the mean diversity obtained by the two aggregation strategies for groups with size = 5

Compared to the strategy that aggregates the recommendations, the aggregating preferences strategy combines the opinions of the different members in a very early stage of the recommendation process, thereby increasing the diversity of the group recommendations for SVD, UBCF and CB. Combining the profiles of the different members leads to a broader group profile containing more items (SVD), which can be linked to more unconsumed items (CB), and to more neighbouring users (UBCF). However since the group ratings are an average of the members' ratings, the group ratings are less extreme (i.e., closer to the middle point of the rating scale). Since the IBCF suggests the items that are most similar to the highest rated items in the profile, the recommendations based on IBCF are less diverse if the aggregating preferences strategy is used.

4.3.2.4 Coverage influenced by the group size

Figure 4.6 shows the mean coverage of the recommendations (calculated over all users) together with the 95% confidence interval of the mean coverage, in relation to the recommendation algorithm, aggregation strategy, and the group size. Since the most-popular recommender always suggests the same list of items for all users or groups regardless the preferences of the users or the size of the group, the coverage of this recommender is very low (i.e., $5/1682 = 0.00297$). Therefore, the horizontal axis crosses the vertical axis at the origin.

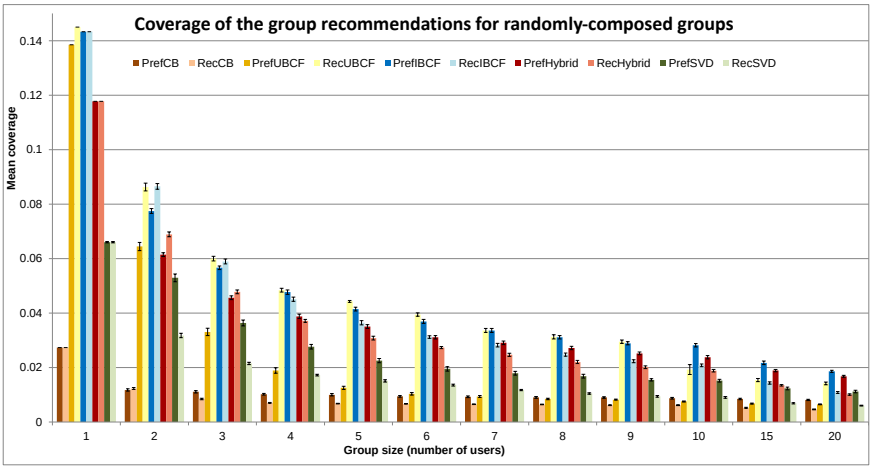


Figure 4.6: The coverage of the group recommendations for randomly-composed groups of a varying group size

The CB recommender has the lowest catalog coverage. Because these recommendations are merely based on the metadata of the items, different groups often

receive suggestions for the same items. The coverage of the recommender based on SVD is considerably higher. The recommendation lists generated by UBCF and IBCF have the least overlap for the different groups and as a result, these algorithms have the highest coverage. The coverage of the Hybrid recommender is mainly due to the high coverage of the CF algorithm.

As expected, Figure 4.6 shows for all algorithms a decreasing coverage when the group size increases. Since all users are a member of only one group (as specified in Section 4.2.3.1), the number of groups decreases as the group size increases. So, more users are combined in a single group and all members of the group receive the same group recommendations. Consequently, as the group size increases, more users receive the same group recommendations and as a result the coverage decreases.

For most algorithms, the coverage obtained by using the aggregating preferences strategy is slightly higher than the coverage of the aggregated recommendations. One exception is UBCF, which has a higher catalog coverage in combination with the aggregating recommendations strategy than with the aggregating preferences strategy. Table 4.3 shows the results of the statistical T-tests comparing the mean coverage of the recommendations generated by the two aggregation strategies for groups of five members. (Similar results are obtained for other group sizes.) H_0 = the mean coverage of the recommendations generated by the aggregating preferences strategy is equal to the mean coverage of the recommendations generated by the aggregating recommendations strategy. T-tests with a p-value below 0.05 indicate a significant difference between the two aggregation strategies, and are presented in bold.

Algorithm	t(58)	p-value
CB	12.36	**0.00**
UBCF	-81.64	**0.00**
IBCF	8.16	**0.00**
Hybrid	7.29	**0.00**
SVD	15.51	**0.00**

Table 4.3: Statistical T-test comparing the mean coverage obtained by the two aggregation strategies for groups with size = 5

4.3.2.5 Serendipity influenced by the group size

Figure 4.7 shows the mean serendipity of the recommendations (calculated over all users) together with the 95% confidence interval of the mean serendipity, in relation to the recommendation algorithm, aggregation strategy, and the group size.

The serendipity value of the list of popular recommendations is 0.43, which is indicated in Figure 4.7 by the level of the horizontal axis. Since the popular recommendations are based on the consumption behaviour of the whole community, this recommendation list might contain items that are unknown for some users and thereby seem surprising. E.g., the list can contain movies of a genre that the user has never watched before. So in general, the list of most popular items is rather serendipitous for the users.

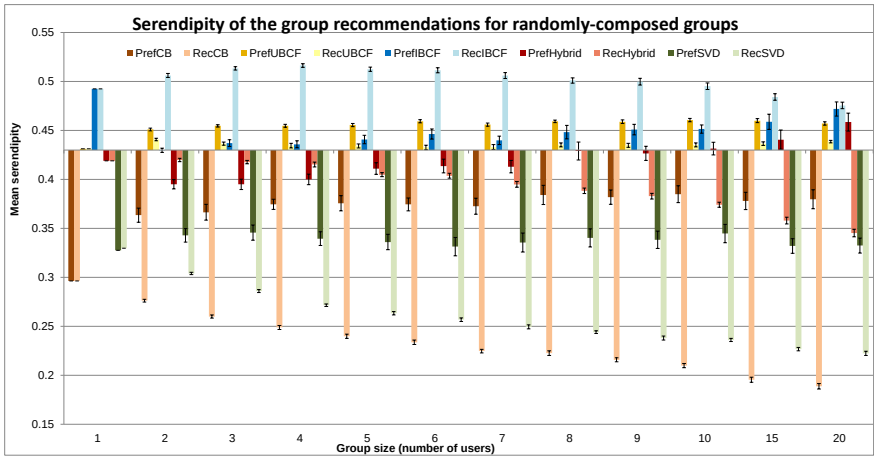


Figure 4.7: The serendipity of the group recommendations for randomly-composed groups of a varying group size

In contrast, the recommendation lists of the SVD and CB recommender contain items that users may expect. These recommenders mainly suggest items of the same genres as the items in the profile of the user, thereby not surprising the user. Consequently, the serendipity of the SVD and CB recommender is significantly lower than the serendipity of the most-popular recommender. Also the Hybrid recommender suffers from these 'too obvious' recommendations of the CB recommender. On the other hand, algorithms based on CF have the potential for serendipitous recommendations, which might be more interesting, surprising, and useful for the users.

The serendipity of most algorithms' recommendations remains constant as the group size increases. As with the diversity of the recommendations, RecCB, RecSVD, and RecHybrid are the only exceptions, showing a decreased serendipity as the group size increases.

Comparing the two aggregation strategies shows that SVD, UBCF, and the CB recommender produce the most serendipitous recommendations if the preferences are aggregated whereas the group recommendations of IBCF are more serendip-

itous if the members' individual recommendation lists are aggregated. For the Hybrid recommender, the aggregation strategy that leads to the most serendipitous recommendations depends on the group size. Table 4.4 shows the results of the statistical T-tests comparing the mean serendipity of the recommendations generated by the two aggregation strategies for groups of five members. (Similar results are obtained for other group sizes.) H_0 = the mean serendipity of the recommendations generated by the aggregating preferences strategy is equal to the mean serendipity of the recommendations generated by the aggregating recommendations strategy. T-tests with a p-value below 0.05 indicate a significant difference between the two aggregation strategies, and are presented in bold.

Algorithm	t(58)	p-value
CB	28.59	**0.00**
UBCF	13.72	**0.00**
IBCF	-25.18	**0.00**
Hybrid	1.69	0.10
SVD	15.31	**0.00**

Table 4.4: Statistical T-test comparing the mean serendipity obtained by the two aggregation strategies for groups with size = 5

4.3.3 Influence of the intra-group similarity

4.3.3.1 Intra-group similarity experiment

The third series of experiments (Section 4.3.3.2, 4.3.3.3, 4.3.3.4, 4.3.3.5) investigates the influence of the similarity of group members on the quality of the group recommendations. In this series of experiments, the groups are composed of users which are more or less similar to each other.

For each measurement, the groups are created as follows. First a *minimum intra-group similarity* is determined. This is a minimum threshold for the similarity of each couple of members in the group. So each couple of users of the same group needs to have a user-user similarity that is equal to or greater than this minimum intra-group similarity. These user-user similarities are calculated by using the Pearson correlation metric on the users' ratings in the data set.

Then, groups are composed by selecting users who fulfil the requirement of the minimum intra-group similarity. The first member of the group is randomly selected without any requirement; the second member is randomly selected from the subset of users who are sufficiently similar to the first user. So the second user has a user-user similarity with the first user which is at least the defined minimum intra-group similarity. The third member of the group is randomly selected from

the subset of users who are sufficiently similar to the first and the second user. This process of adding similar users to the group is repeated until the intended group size is reached. Each user can be selected for only one group, in which (s)he meets the requirement of the intra-group similarity. The result is a group of users in which every user is similar to every other user of the group with a minimum similarity as defined by the minimum intra-group similarity.

To investigate the influence of the intra-group similarity separately, the group size is fixed in these experiments whereas the minimum intra-group similarity is varying from -1.00 to 0.80 if the group size is 2, and from -1.00 to 0.55 if the group size is 5. Only the results for groups of 2 members (in Figures 4.8, 4.10, 4.11, and 4.12) and 5 members (in Figure 4.9) are included in this dissertation, since the graphs for other group sizes result in similar findings.

The minimum intra-group similarity starts at -1.00, i.e., the lowest similarity value that can be obtained by using the Pearson correlation metric. This minimum intra-group similarity of -1.00 denotes that all users are a candidate to be combined into a group. Group members can have similar preferences but they can also have completely opposite preferences. This situation corresponds to the random group composition of Section 4.3.2 in which no restrictions are imposed on the group.

Further, the quality of the group recommendations is evaluated for groups with a minimum intra-group similarity of -0.75, -0.50 and -0.25. This means that the members can still have conflicting preferences but users who are complete opposites of each other (similarity of -1.00) are not allowed in the same group. Groups with a minimum intra-group similarity of 0.00 consist only of users with non-conflicting preferences; i.e., the user-user similarity of each couple of members is always positive. From then on, the recommendations are evaluated for groups with a minimum intra-group similarity that varies in steps of 0.05. As the minimum intra-group similarity increases, the condition for a user to join a group is becoming stricter. Group members have to be more similar to each other and the group becomes a homogeneous set of like-minded users.

For a group size of 2, the process of combining more similar users is stopped at a minimum intra-group similarity of 0.80. For higher values of the minimum intra-group similarity, it is not possible any more to find a sufficient number of groups in which all users are so similar to each other. For groups of 5 users, it is even more difficult to find members who are all very similar to each other. Therefore, the minimum intra-group similarity is increased until 0.55 is reached.

Given the random aspect in the group composition (i.e., selecting a random user from the subset of users who are sufficiently similar to the other group members), the process of composing groups is repeated 30 times. Similar to the procedure of the first and second series of experiments, each quality metric is calculated

30 times and the mean of these measurements is used as an estimation of the quality of the group recommendations.

So the graphs in Figures 4.8, 4.9, 4.10, 4.11, and 4.12 show these mean values, as well as the 95% confidence intervals of the mean values. Again, the vertical axis of each figure crosses the horizontal axis at the quality level of the most-popular recommender and the prefix of the bar series denotes if the algorithm uses the aggregating recommendations strategy ("Rec") or the aggregating preferences strategy ("Pref"). Also in these experiments the average function is used as aggregation method to combine the individual preferences or recommendation lists.

4.3.3.2 Accuracy influenced by the intra-group similarity

Figure 4.8 shows the mean nDCG (calculated over all users) for groups of two members together with the 95% confidence interval of the mean nDCG, in relation to the recommendation algorithm, aggregation strategy, and the minimum intra-group similarity. In this graph, two horizontal lines are indicating the accuracy of recommendations that are calculated for individual users. The green line (bottom line) represents the accuracy of recommendations calculated by the CB algorithm; this recommender has the lowest accuracy score for individual users. The red line (upper line) indicates the highest accuracy level that was obtained for individual recommendations; these recommendations are generated using SVD.

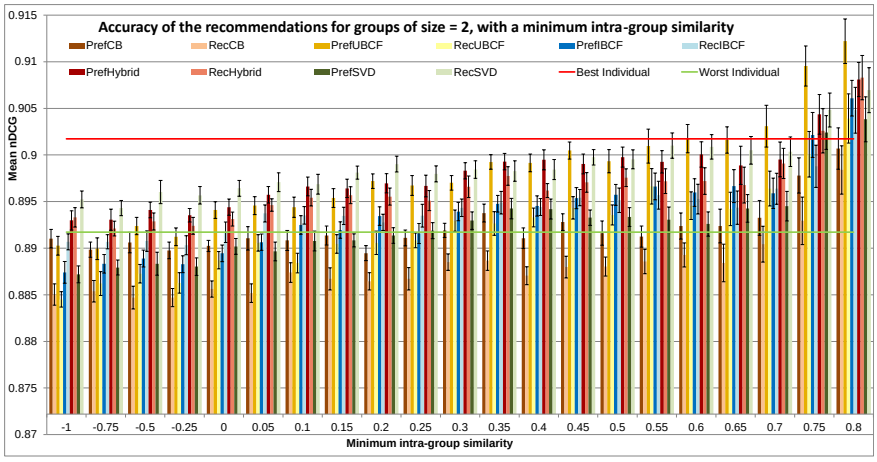


Figure 4.8: The accuracy of the group recommendations for groups of size = 2, with a minimum intra-group similarity

As was already discovered by Baltrunas et al. [1], the accuracy of the group recommendations increases as the similarity between members of the group increases. The more similar the members of the group, the higher the accuracy of

the group recommendations. This accuracy difference is especially noticeable for groups with a high intra-group similarity. If the minimum intra-group similarity is 0.60, the recommendations for groups of two members generated by UBCF are about as accurate as the most accurate recommendations for individuals (generated using SVD). For higher values of the minimum intra-group similarity, the accuracy of the group recommendations can transcend the accuracy level of recommendations for individuals. For example, if the minimum intra-group similarity is 0.80, all algorithms, except for the CB recommender, generate group recommendations that have a higher accuracy than the most-accurate recommendations for individuals.

This effect is even more pronounced for larger groups. Figure 4.9 shows the mean nDCG (calculated over all users) for groups of five members together with the 95% confidence interval of the mean nDCG, in relation to the recommendation algorithm, aggregation strategy, and the minimum intra-group similarity. In comparison with the results of Figure 4.8, the accuracy of the recommendations for groups of five members is increasing faster as the minimum intra-group similarity increases. As soon as the minimum intra-group similarity is 0.25, the accuracy level of recommendations for individuals is reached. For groups of very similar users, the group recommendations of all algorithms show a significantly increased accuracy, thereby outperforming the recommendations for individuals. So if similar users are brought together in groups, even the least accurate algorithm (CB) can generate group recommendations that are more effective than the best recommendations calculated for each individual seperately.

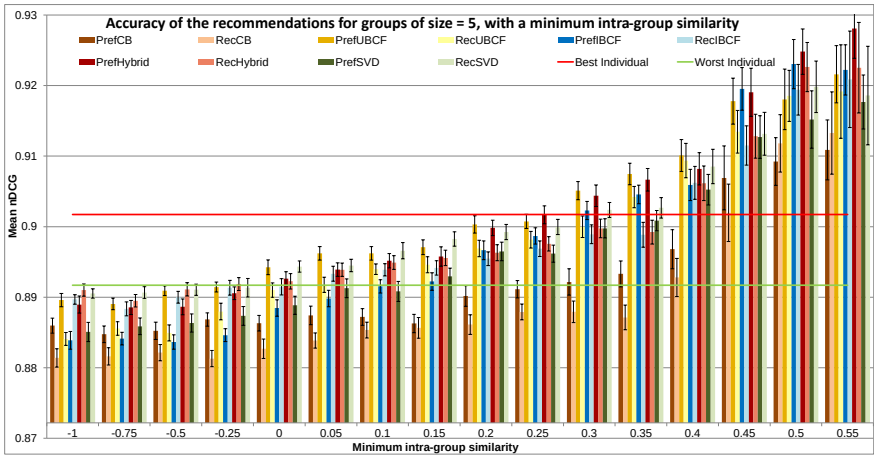


Figure 4.9: The accuracy of the group recommendations for groups of size = 5, with a minimum intra-group similarity

Important to keep in mind is the fact that Figures 4.8 and 4.9 show the mean nDCG for each value of the minimum intra-group similarity. So for some users the recommendations based on their individual preferences are most accurate, whereas for other users their group recommendations based on the preferences of all group members' are most accurate.

If groups are randomly composed, group members may have different or even conflicting preferences. Group recommenders have then the challenging task to generate suggestions that please all group members. Since it is not always possible to find items perfectly matching the tastes of all members, the accuracy of the group recommendations might be lower than the accuracy of the recommendations based on the individual preferences.

In contrast, if groups are composed of users with similar preferences, group recommenders do not have to deal with conflicting preferences and items that match each group member's tastes can easily be found. Moreover, the group members are complementary to each other and can learn from each other's experiences with previously consumed content. If group members are similar, they will often have a comparable rating behaviour. Thus, the rating of one member can be a good estimation of the rating of another member for the same item. As a results, one member's ratings can enrich the profile of another member since the ratings of both users are highly correlated. The more similar the members, the better they can complement each other, resulting in more accurate recommendations, as shown in Figures 4.8 and 4.9. Table 4.5 confirms this by the results of the statistical T-tests comparing the mean accuracy of the recommendations for groups of two members (size = 2) with a minimum intra-group similarity of -1.0 and 0.5. (Similar results are obtained for other group sizes.) H_0 = the mean accuracy of the recommendations generated for groups with a minimum intra-group similarity of -1.0 is equal to the mean accuracy of the recommendations generated for groups with a minimum intra-group similarity of 0.5. T-tests with a p-value below 0.05 indicate a significant difference between the two values of the minimum intra-group similarity, and are presented in bold.

So compared to randomly-composed groups, a significant accuracy improvement is obtained for all algorithms (except for PrefCB this improvement was not significant) when the group members are similar to each other. Since the accuracy gain obtained by the similarity of group members is varying for each group, the standard deviation of the accuracy slightly increases as the minimum intra-group similarity increases. This is indicated by the size of the confidence intervals in Figures 4.8 and 4.9.

Besides the similarity of the group members, the size of the group has also an influence on the accuracy. The comparison of Figures 4.8 and 4.9 shows that if groups are randomly composed (minimum intra-group similarity of -1.00), group

Algorithm	t(58)	p-value
PrefCB	-0.62	0.53
PrefUBCF	-9.64	**0.00**
PrefIBCF	-8.74	**0.00**
PrefHybrid	-7.76	**0.00**
PrefSVD	-5.27	**0.00**
RecCB	-3.35	**0.00**
RecUBCF	-11.33	**0.00**
RecIBCF	-4.83	**0.00**
RecHybrid	-5.13	**0.00**
RecSVD	-5.38	**0.00**

Table 4.5: Statistical T-test comparing the mean accuracy obtained for groups of two members (group size = 2) with a minimum intra-group similarity of -1.0 and 0.5

recommendations are most accurate for small groups. In contrast, if members are similar to each other, larger groups (Figure 4.9) can lead to more accurate group recommendations than smaller groups (Figure 4.8). E.g., the recommendations for a group of five members with a minimum intra-group similarity of 0.50 have a significantly higher accuracy than the recommendations for a group of two members with the same minimum intra-group similarity. The more users in a group, the more information and preferences that can be shared among group members. So, if these group members are similar to each other, larger groups can result in more accurate group recommendations.

4.3.3.3 Diversity influenced by the intra-group similarity

Figure 4.10 shows the mean list diversity (calculated over all users) for groups of two members together with the 95% confidence interval of the mean list diversity, in relation to the recommendation algorithm, aggregation strategy, and the minimum intra-group similarity.

The results show that for PrefUBCF the list diversity slightly decreases as the minimum intra-group similarity increases. If group members are very similar to each other, all members have the same or very similar items in their profile. Aggregating these individual profiles leads to little variety in the group profile. Consequently, the recommended items are very similar to each other and so the list diversity decreases as the minimum intra-group similarity increases.

For PrefSVD and PrefIBCF on the other hand, the list diversity slightly increases as the minimum intra-group similarity increases. In contrast to UBCF, SVD and IBCF do not create a user profile modelling the user's preferences in order to generate recommendations. The increasing diversity of the PrefHybrid

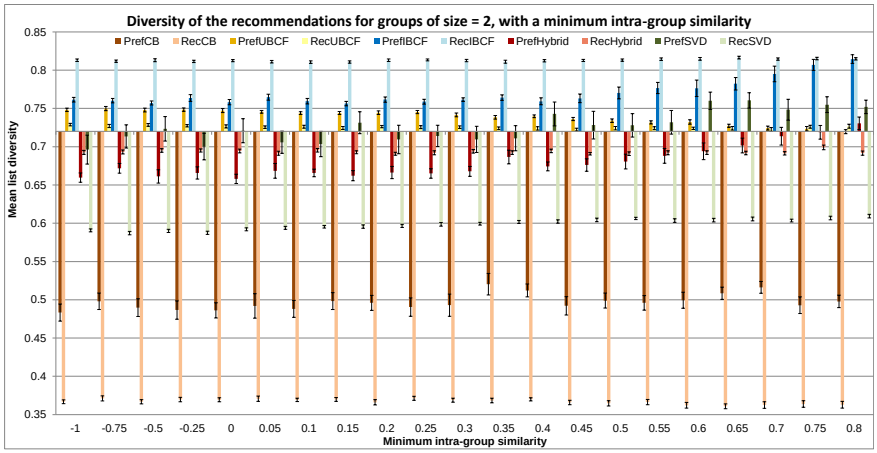


Figure 4.10: The diversity of the group recommendations for groups of size = 2, with a minimum intra-group similarity

algorithm is due to the increasing diversity of the underling IBCF algorithm. Also for RecSVD a slight increase in diversity is witnessed.

For the other algorithms, the list diversity remains constant as the minimum intra-group similarity increases, meaning that the similarity between group members has no influence on the list diversity.

Table 4.6 shows the results of the statistical T-tests comparing the mean diversity of the recommendations for groups of two members (size = 2) with a minimum intra-group similarity of -1.0 and 0.5. (Similar results are obtained for other group sizes.) H_0 = the mean diversity of the recommendations generated for groups with a minimum intra-group similarity of -1.0 is equal to the mean diversity of the recommendations generated for groups with a minimum intra-group similarity of 0.5. T-tests with a p-value below 0.05 indicate a significant difference between the two values of the minimum intra-group similarity, and are presented in bold.

4.3.3.4 Coverage influenced by the intra-group similarity

Figure 4.11 shows the mean coverage of the recommendations (calculated over all users) for groups of two members together with the 95% confidence interval of the mean coverage, in relation to the recommendation algorithm, aggregation strategy, and the minimum intra-group similarity.

The catalog coverage generally remains constant as the minimum intra-group similarity increases. So, the similarity between group members has no noteworthy influence on the catalog coverage of the group recommendations. An exception is

Algorithm	t(58)	p-value
PrefCB	1.79	0.08
PrefUBCF	13.15	**0.00**
PrefIBCF	-1.81	0.07
PrefHybrid	-3.13	**0.00**
PrefSVD	-2.22	**0.03**
RecCB	0.76	0.45
RecUBCF	3.00	**0.00**
RecIBCF	-0.17	0.87
RecHybrid	0.86	0.40
RecSVD	-10.74	**0.00**

Table 4.6: Statistical T-test comparing the mean diversity obtained for groups of two members (group size = 2) with a minimum intra-group similarity of -1.0 and 0.5

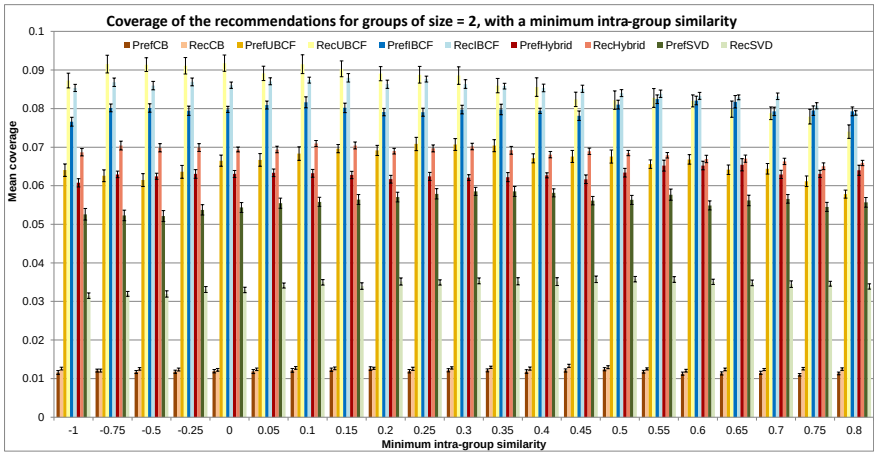


Figure 4.11: The coverage of the group recommendations for groups of size = 2, with a minimum intra-group similarity

the coverage of RecUBCF and RecIBCF that slightly decreases as the minimum intra-group similarity increases. So, these algorithms have the highest coverage for randomly-composed groups (minimum intra-group similarity = -1.00), but this coverage may be slightly lower if the group members are more similar to each other.

Table 4.7 shows the results of the statistical T-tests comparing the mean coverage of the recommendations for groups of two members (size = 2) with a minimum intra-group similarity of -1.0 and 0.5. (Similar results are obtained for other group sizes.) H_0 = the mean coverage of the recommendations generated for groups with

a minimum intra-group similarity of -1.0 is equal to the mean coverage of the recommendations generated for groups with a minimum intra-group similarity of 0.5. T-tests with a p-value below 0.05 indicate a significant difference between the two values of the minimum intra-group similarity, and are presented in bold.

Algorithm	t(58)	p-value
PrefCB	-2.22	**0.03**
PrefUBCF	-2.57	**0.01**
PrefIBCF	-4.61	**0.00**
PrefHybrid	-2.91	**0.01**
PrefSVD	-3.34	**0.00**
RecCB	-1.08	0.28
RecUBCF	4.82	**0.00**
RecIBCF	1.77	0.08
RecHybrid	0.32	0.75
RecSVD	-7.08	**0.00**

Table 4.7: Statistical T-test comparing the mean coverage obtained for groups of two members (group size = 2) with a minimum intra-group similarity of -1.0 and 0.5

Although Table 4.7 identifies significant difference for multiple algorithms, Figure 4.11 shows that these differences in coverage are mostly small, and that a trend on the basis of the minimum intra-group similarity is often missing.

4.3.3.5 Serendipity influenced by the intra-group similarity

Figure 4.12 shows the mean serendipity of the recommendations (calculated over all users) for groups of two members together with the 95% confidence interval of the mean serendipity, in relation to the recommendation algorithm, aggregation strategy, and the minimum intra-group similarity.

For PrefSVD and PrefIBCF (and to a lesser extent for RecSVD) the serendipity increases as the minimum intra-group similarity increases. These findings are in accordance with the results for PrefSVD, PrefIBCF, and RecSVD of Section 4.3.3.3, which show an increased list diversity for similar group members. So, if recommendations are more diverse, they are probably more serendipitous for the user. The serendipity of the recommendations generated by other algorithms remains constant as the minimum intra-group similarity increases.

Table 4.8 shows the results of the statistical T-tests comparing the mean serendipity of the recommendations for groups of two members (size = 2) with a minimum intra-group similarity of -1.0 and 0.5. H_0 = the mean serendipity of the recommendations generated for groups with a minimum intra-group similarity of -1.0 is

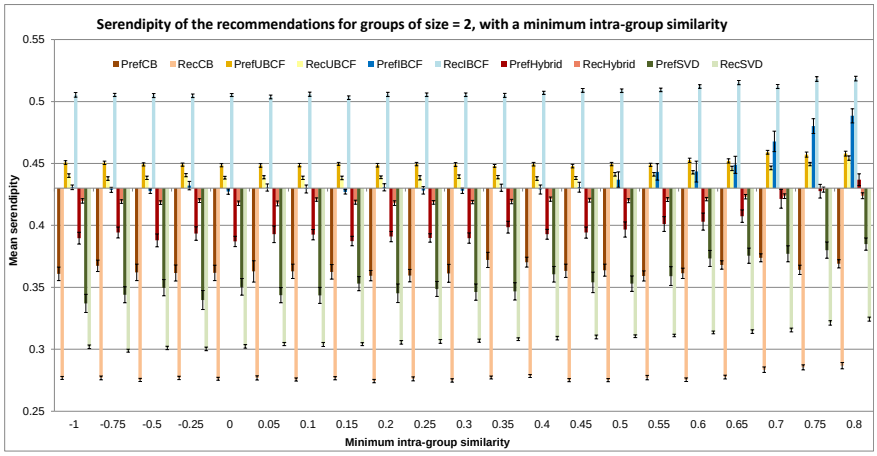


Figure 4.12: The serendipity of the group recommendations for groups of size = 2, with a minimum intra-group similarity

equal to the mean serendipity of the recommendations generated for groups with a minimum intra-group similarity of 0.5. T-tests with a p-value below 0.05 indicate a significant difference between the two values of the minimum intra-group similarity, and are presented in bold.

Algorithm	t(58)	p-value
PrefCB	-0.68	0.50
PrefUBCF	7.95	**0.00**
PrefIBCF	-55.24	**0.00**
PrefHybrid	-1.53	0.13
PrefSVD	-2.83	**0.01**
RecCB	1.61	0.11
RecUBCF	-0.83	0.41
RecIBCF	-2.41	**0.02**
RecHybrid	-0.12	0.90
RecSVD	-8.11	**0.00**

Table 4.8: Statistical T-test comparing the mean serendipity obtained for groups of two members (group size = 2) with a minimum intra-group similarity of -1.0 and 0.5

4.3.4 Improved aggregation strategy

4.3.4.1 Combining strategies

The results of Section 4.3.2.2 showed that the used aggregation strategy in combination with the recommendation algorithm has a major influence on the accuracy of the group recommendations. Certain algorithms (such as CB and UBCF) produce more accurate group recommendations when the aggregating preferences strategy is used, whereas other algorithms (such as IBCF and SVD) obtain a higher accuracy in combination with the aggregating recommendations strategy. So, the choice of the aggregation strategy is crucial for each algorithm in order to obtain the best group recommendations.

Instead of selecting one individual aggregation strategy, traditional aggregation strategies can be combined with the aim of obtaining group recommendations which outperform the group recommendations of each individual aggregation strategy. In this context, Berkovsky and Freyne [13] witnessed that the aggregating recommendations strategy outperforms the aggregating preferences strategy in terms of accuracy if the user profiles have a low density (i.e., containing a low number of consumptions). For these users, of whom little is known from their low-density profile, they obtained the lowest MAE (Mean Absolute Error for the prediction score of the group recommendations) when the aggregating recommendations strategy is used. In contrast for high-density profiles, the aggregating preferences strategy resulted in the lowest MAE, thereby outperforming the aggregating recommendations strategy in terms of accuracy. Therefore, Berkovsky and Freyne proposed a switching scheme that uses the aggregating recommendations strategy in combination with a low-density profile and switches to the aggregating preferences strategy when the user profile becomes denser. Compared to the individual aggregation strategies, this switching strategy yielded a small accuracy improvement.

Inspired by the proposed strategy of Berkovsky and Freyne, we employed a switching scheme that selects either the aggregating preferences strategy or the aggregating recommendations strategy to calculate group recommendations for users of the MovieLens data set. We experimented with various switching thresholds based on the user profile density as well as based on the group profile density. In addition, switching based on the intra-group similarity, i.e., the similarity between group members, was evaluated. However, the group recommendations obtained by using such a switching scheme did not outperform the group recommendations that are based on the best individual aggregation strategy in terms of accuracy. The reason why we could not reproduce the accuracy gain of the switching scheme of Berkovsky and Freyne on the MovieLens data set might be the specific settings of their experiment. They only considered the accuracy of recommendations generated by a CF algorithm, the MAE metric was used to estimate the accuracy, and

they focused on the specific use case of recipe recommendations using a rather small data set (around 3300 ratings).

Therefore, we continued our quest to a more advanced aggregation strategy which combines individual aggregation strategies thereby yielding an accuracy gain compared to each individual aggregation strategy. The aim of this combination of strategies is to merge the knowledge of two (or more) aggregation strategies into a final group recommendation list. The idea is that if one of the aggregation strategies comes up with a less suitable or undesirable group recommendation, the other aggregation strategy can correct this mistake. This makes the group recommendations resulting from the combination of strategies more robust than the group recommendations based on a single aggregation strategy.

Although the aggregation strategies can be combined in various possible ways, our experiments showed that most combination techniques do not obtain an increased accuracy of the group recommendations. According to the results of our experiments, an effective way to generate group recommendations by combining the two aggregation strategies is as follows: First, group recommendations are calculated by using the selected recommendation algorithm and the aggregating preferences strategy. The result is a list of all items, ordered according to their prediction score, which estimates how much each item will be appreciated by the group. In case of an individual aggregation strategy, the top-N items on that list are selected as suggestions for the group. After calculating the group recommendations using the aggregating preferences strategy, or in parallel with it, group recommendations are generated using the chosen algorithm and the aggregating recommendations strategy. Again, the result is an ordered list of items with their corresponding prediction score.

Subsequently, the two item lists are combined into one item list by combining the prediction scores of each aggregation strategy per item. In this experiment, we opted for the average (arithmetic mean) as method to combine the prediction scores. So in the resulting item list, each item's prediction score is the average of the item's prediction score generated by the aggregating preferences strategy and the item's prediction score produced by the aggregating recommendations strategy. Alternative combining methods are also possible, e.g., a weighted average of the prediction scores with weights depending on the performance of each individual aggregation strategy. Then the items are ordered by their new prediction score in order to obtain a new combined list of potential group recommendations.

This combined item list can still contain items that are at the top of the recommendation list that is generated by one of the aggregation strategies but that are in the middle or even at the bottom of the recommendation list produced by using the other aggregation strategy. Therefore, the combined item list is adapted in order to contain only items that appear at the top of both recommendation lists, thereby reducing the risk of recommending undesirable or less suitable items to

the group. So, items that are ranked below a certain threshold position in the recommendation list generated by one of the aggregation strategies, are removed from the combined list. In this experiment, we opted to exclude these items from the combined list, that are not in the top-5% of both recommendation lists (i.e., the top-84 of recommended items for the MovieLens data set). Since only a limited number of recommendations are offered to the users, (5 in our experiment,) the filtering of the top-5% items is no hard restriction. As a result, the final recommendation list contains the items that are identified as 'the most suitable' by both aggregation strategies, ordered according to the average of the prediction scores of both aggregation strategies.

4.3.4.2 Accuracy improvement by combining strategies

Our combined aggregation strategy is compared to the individual aggregation strategies in Figure 4.13. Similar to the experiments of Section 4.3.2, the groups are composed by selecting random users from the data set and the process of composing groups is repeated 30 times. So, the graph of Figure 4.13 shows the mean accuracy of these measurements as an estimation of the quality of the group recommendations (on the vertical axis), as well as the 95% confidence interval of the mean value, in relation to the recommendation algorithm, aggregation strategy, and the group size. The group size is indicated on the horizonal axis. Again, the vertical axis of each figure crosses the horizontal axis at the quality level of the most-popular recommender and the prefix of the bar series denotes which aggregation strategy is used. The prefix ("Combined") stands for the proposed aggregation strategy which combines the aggregating preferences strategy and the aggregating recommendations strategy. The bar series with the prefix ("Best") indicates the accuracy of the best individual strategy, i.e., aggregating preferences for the UBCF and CB algorithm, and aggregating recommendations for the SVD, IBCF, and Hybrid algorithm. For the individual aggregation strategies, the average (Avg) function is used as aggregation method to combine the individual preferences or recommendations.

The non-overlapping confidence intervals indicate a significant improvement of the combined aggregation strategy compared to the best individual aggregation strategy. Table 4.9 shows the results of the statistical T-tests comparing the mean accuracy of the recommendations generated by the best individual aggregation strategy and by the combined aggregation strategy for groups with size = 5. (Similar results are obtained for other group sizes.) H_0 = the mean accuracy of the recommendations generated by using the best individual aggregation strategy is equal to the accuracy of the recommendations generated by using the combined aggregation strategy. The small p-values (all smaller than 0.05) prove the significant accuracy improvement of our proposed aggregation strategy. However, this

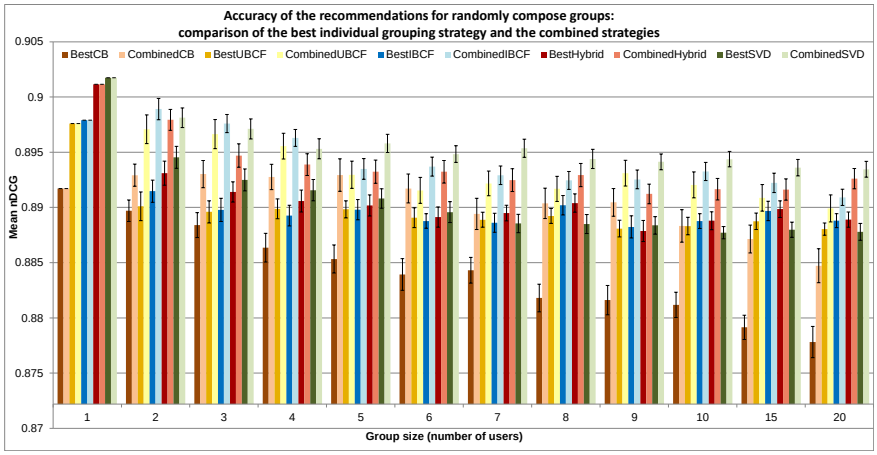


Figure 4.13: The accuracy of the group recommendations for randomly-composed groups of a varying group size using the best individual aggregation strategy and the combined aggregation strategy

combined aggregation strategy has also a disadvantage. Since it uses the output of the individual aggregation strategies, group recommendations have to be calculated for each individual strategy (two times in this experiment). As a result, the calculation load increases linearly with the number of aggregation strategies that have to be combined. Fortunately, these calculations can be parallelized to speed up the total computation time.

Algorithm	t(58)	p-value
CB	-3.55	**0.00**
UBCF	-2.66	**0.01**
IBCF	-2.33	**0.02**
Hybrid	-2.53	**0.01**
SVD	-4.39	**0.00**

Table 4.9: Statistical T-test comparing the accuracy obtained by using the best individual aggregation strategy and the combined aggregation strategy for groups with size = 5

4.4 Conclusions

In this chapter, group recommendations are discussed as a solution to generate suggestions for a group of people (such as a family or friends). A typical use-case is a recommender system for audio and video content in the home environment. Multiple qualitative aspects (accuracy, diversity, coverage, and serendipity) of the group recommendations are thoroughly evaluated for five state-of-the-art recommendation algorithms in combination with two commonly-used aggregation strategies. Furthermore, the influence of the group size and group composition on the effectiveness of the group recommendations is investigated.

The results of this chapter are summarized per section in Table 4.10. An important result is the finding that there exists no 'overall-best' recommendation algorithm and aggregation strategy. The recommendation algorithm and aggregation strategy should be chosen together in order to optimize the desired qualitative aspects of the group recommendations. E.g., if the main objective of the group recommender system is to achieve a high accuracy for small to medium sized groups (size < 7), we recommend using the SVD algorithm in combination with the aggregating recommendations strategy. If other quality aspects such as diversity or coverage are also important, we recommend the IBCF or Hybrid algorithm with the aggregating recommendations strategy. When a recommender system for individual users is extended to enable group recommendations, these results can be used to choose the best aggregation strategy based on the currently employed algorithm.

Future research can include the evaluation of the effectiveness of the group recommendations via an online experiment with real test subjects. In such an experiment, users can be invited to use the group recommender system at home with their family and evaluate the group recommendations afterwards. An online experiment makes it possible to investigate if the results of the offline analysis are in line with the assessments of the users and if differences in accuracy, diversity, and serendipity are noticeable for these users.

References

[1] L. Baltrunas, T. Makcinskas, and F. Ricci. *Group recommendations with rank aggregation and collaborative filtering*. In Proceedings of the fourth ACM conference on Recommender systems, RecSys '10, pages 119–126, New York, NY, USA, 2010. ACM.

[2] Y.-L. Chen, L.-C. Cheng, and C.-N. Chuang. *A group recommendation system with consideration of interactions among group members*. Expert Systems with Applications, 34(3):2082–2090, 2008.

[3] L. Quijano-Sanchez, J. A. Recio-Garcia, and B. Diaz-Agudo. *Personality and Social Trust in Group Recommendations*. In Proceedings of the 2010 22nd IEEE International Conference on Tools with Artificial Intelligence - Volume 02, ICTAI '10, pages 121–126, Washington, DC, USA, 2010. IEEE Computer Society.

[4] Grouplens Research. *MovieLens Data Sets*, 2011. Online available at `http://www.grouplens.org/node/73`.

[5] J. Masthoff. *Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers*. User Modeling and User-Adapted Interaction, 14:37–85, 2004.

[6] M. Deshpande and G. Karypis. *Item-based top-N recommendation algorithms*. ACM Transactions on Information Systems, 22(1):143–177, 2004.

[7] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. *Improving recommendation lists through topic diversification*. In Proceedings of the 14th international conference on World Wide Web, WWW '05, pages 22–32, New York, NY, USA, 2005. ACM.

[8] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. *Evaluating collaborative filtering recommender systems*. ACM Transactions on Information Systems, 22(1):5–53, 2004.

[9] M. Ge, C. Delgado-Battenfeld, and D. Jannach. *Beyond accuracy: evaluating recommender systems by coverage and serendipity*. In Proceedings of the fourth ACM conference on Recommender systems, RecSys '10, pages 257–260, New York, NY, USA, 2010. ACM.

[10] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor. *Recommender Systems Handbook*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

[11] T. Murakami, K. Mori, and R. Orihara. *Metrics for Evaluating the Serendipity of Recommendation Lists*. In K. Satoh, A. Inokuchi, K. Nagao, and T. Kawamura, editors, New Frontiers in Artificial Intelligence, volume 4914 of *Lecture Notes in Computer Science*, pages 40–46. Springer Berlin Heidelberg, 2008.

[12] A. Jameson and B. Smyth. *The adaptive web*. chapter Recommendation to groups, pages 596–627. Springer-Verlag Berlin Heidelberg, 2007.

[13] S. Berkovsky and J. Freyne. *Group-based recipe recommendations: analysis of data aggregation strategies*. In Proceedings of the fourth ACM conference on Recommender systems, RecSys '10, pages 111–118, New York, NY, USA, 2010. ACM.

Experiment	Section	Results
Aggregation method	4.3.1.3	The 'average' and 'average without misery' method generally produce the most accurate group recommendations. The 'one user choice' method induces a low accuracy in combination with the aggregating recommendations strategy.
Accuracy random groups	4.3.2.2	The accuracy of the recommendations decreases as the group size increases. The aggregation strategy that generates the most accurate group recommendations depends on the algorithm. For CB and UBCF, aggregating preferences is the best strategy. For SVD and IBCF, the best strategy is aggregating the recommendations.
Diversity random groups	4.3.2.3	The CB algorithm generates the least diverse recommendation list, even less diverse than the most-popular list. Algorithms based on CF generate the most diverse recommendations. For most algorithms, the diversity remains constant as the group size increases. For SVD, UBCF, and CB, the aggregating preferences strategy generates the most diverse recommendations. For IBCF, the aggregating recommendations strategy generates the most diverse recommendations.
Coverage random groups	4.3.2.4	The CB recommender has the lowest coverage. Recommenders based on CF have the highest coverage. For most algorithms (except UBCF) and group sizes, the coverage obtained using the aggregating preferences strategy is slightly higher than the coverage of the aggregated recommendations.
Serendipity random groups	4.3.2.5	The serendipity of the recommendations generated by the SVD and CB algorithm is significantly lower than the serendipity obtained by the most-popular recommender. Algorithms based on CF have the potential for serendipitous recommendations. The serendipity of most algorithms' recommendations remains constant as the group size increases. The SVD, UBCF, and CB recommender produce the most serendipitous recommendations if the preferences are aggregated whereas the recommendations of IBCF are most serendipitous if the members' individual recommendation lists are aggregated.
Accuracy similar groups	4.3.3.2	The more similar the members of the group, the higher the accuracy of the recommendations. Compared to randomly-composed groups, the group recommendations show a significantly increased accuracy for groups of similar users, with the potential of outperforming the recommendations for individuals.
Diversity similar groups	4.3.3.3	For most algorithms, the list diversity remains constant as the similarity between group members increases. For PrefSVD, PrefIBCF, and RecSVD on the other hand, the list diversity slightly increases as the similarity between group members increases.
Coverage similar groups	4.3.3.4	The coverage generally remains constant as the similarity between group members increases.
Serendipity similar groups	4.3.3.5	For PrefSVD and PrefIBCF (and to a lesser extent for RecSVD), the serendipity increases as the similarity between group members increases. The serendipity of the recommendations generated by other algorithms remains constant as the similarity between group members increases.
Combining Strategies	4.3.4.2	Compared to the best individual aggregation strategy, a significant accuracy improvement is obtained by combining both strategies.

Table 4.10: Conclusions of the study on group recommendations

Part II

Quality of experience

5 Introduction to quality of experience

5.1 Introduction

Mobile devices are becoming the primary tools for Internet access and communication. According to the latest version of the TNS Mobile Life survey, claimed to be the largest study into mobile consumers [1], this growth in the mobile communication domain is driven by an increased demand for mobile video services. Recent forecasts [2] state that mobile video transmission will generate 70% of the global mobile data traffic by 2016. However, offering a good experience to users remains challenging, and given the dependence on several influencing factors, this especially holds true in the context of mobile video applications [3]. This emphasizes the necessity for service providers to investigate the quality of users' experiences in view of matching the produced video quality to users' subjective expectations.

5.2 QoS vs. QoE

Traditionally, network operators and service providers used to pay close attention to the Quality of Service (QoS). QoS is defined by the International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) as "the collective effect of service performance" [4]. The concept of QoS refers to several related aspects, QoS parameters, which are mutually correlated and all have an influence on the performance of a service. Figure 5.1 lists the QoS parameters that are important for services that enable watching video transmitted over a network.

In the scenario of real-time streaming of multimedia (e.g., voice over IP, online games, and live Internet Protocol TeleVision (IP-TV)), the delay of data packets (sometimes called the latency) is an important QoS parameter that have to be minimized. A measure for this delay is the Round-Trip delay Time (RTT) which is defined as the time it takes for a message to be sent plus the time it takes for an acknowledgement of that message to be received. The variability over time of the packet delay across the network is called jitter. Delay and jitter may disturb the synchronization between different data streams in the network. The network bandwidth expresses the amount of data that can be transmitted over the network. If not enough capacity is available, packet losses may occur. The transport protocol can try to recover from these packet losses by retransmissions, but this is at the expense of an additional delay.

Targets for these QoS parameters have to be defined in association with the parameters of the video source such as the resolution, codec and container, bit rate, frame rate, and aspect ratio. These video parameters have an influence on the network requirements and determine in combination with the QoS parameters the quality of the video at the receiver. E.g., the bit rate of the video source has a direct influence on the required bandwidth of the network; and the effect of packet loss is dependent on the used codec and container for the video. Also the transport mechanism has an impact on the QoS parameters. The network topology and the physical distance between sender and receiver influence the network delay. Faults in the network infrastructure may introduce packet loss, and the transport protocol determines if these lost packets are retransmitted or not.

In the field of computer networking and packet-switched telecommunication networks, QoS is associated with resource reservation control mechanisms rather than the resulting service quality. In this context, QoS refers to the ability to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow. These performance levels are objective measures used to ensure the operation of certain services such as real-time streaming of multimedia. To impose the required performance levels formally, an Service-Level Agreement (SLA) can be defined as part of a service contract. An SLA will typically have a technical definition in terms of Mean Time Between Failures (MTBF), or Mean Time To Repair or Mean Time To Recovery (MTTR). The service provider constantly monitors the network for slow or failing components and intervenes if necessary to ensure the agreed service quality. For the end-user this service quality is noticeable in terms of availability and accessibility of the service.

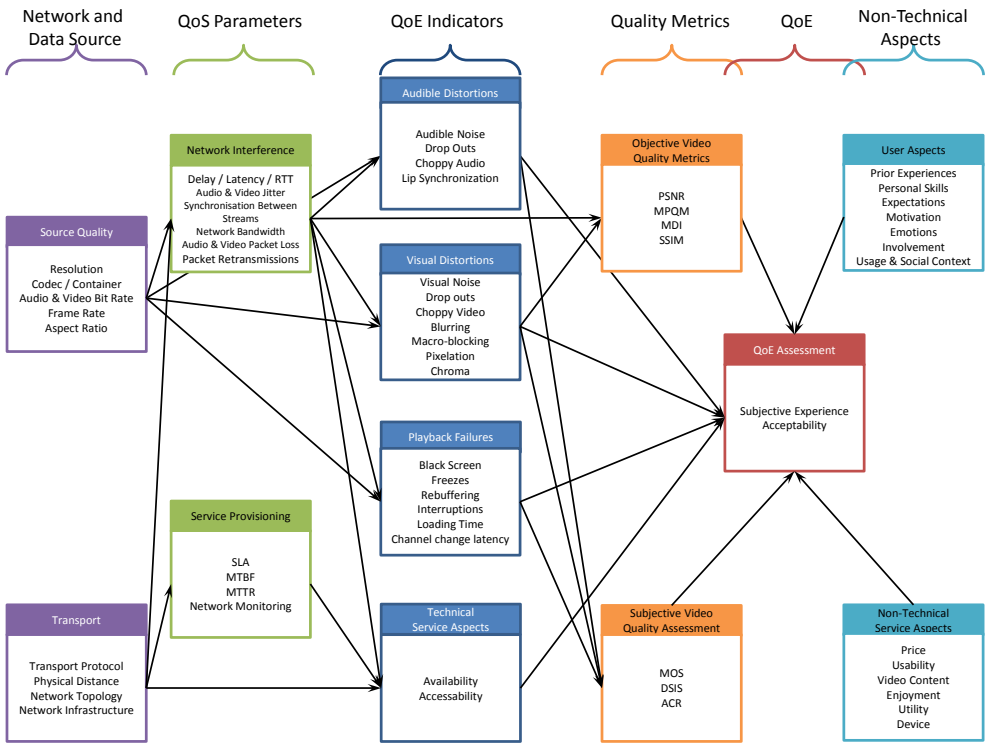


Figure 5.1: Overview of the QoS parameters and their relation with QoE indicators for video watching

Traditionally, the influence of these QoS parameters on the video quality is assessed by objectively-measured video quality metrics. These quality metrics play a crucial role in meeting the promised QoS and in improving the obtained video quality at the receiver side [5]. One of the most commonly used metrics is Peak Signal-to-Noise Ratio (PSNR), which expresses the ratio between the power of the video signal and the power generated by electromagnetic noise in terms of decibels. Besides, various alternative objective quality metrics exists. E.g., Structural SIMilarity (SSIM) is a method for measuring the similarity between two images of a video: the initial distortion-free image and the image that is the result of transmitting the video over the network. This metric uses the structural distortion in video as an estimate of perceived visual distortion. The SSIM metric is designed to improve on traditional methods such as PSNR, which have proved to be inconsistent with human eye perception [6]. The Moving Pictures Quality Metric (MPQM) is used to assess the quality of MPEG compressed video streams. This metric is calculated using a mix of content dependent factors along with a combination of network impairments such as delay and packet loss [7]. Media Delivery Index (MDI) is a metric that indicates video quality levels and also identifies network impairments that are affecting the quality. This is achieved by measuring instances of jitter levels and packet loss occurring on different points in the network [7]. Although this list of metrics is not exhaustive, it indicates that a whole range of different metrics exists for objectively measuring the video quality.

These objective quality assessment methods can be further categorized as full-reference, reduced-reference, and no-reference. Full-reference methods calculate the quality difference by comparing every pixel in each image of the distorted video to its corresponding pixel in the original video. Reduced-reference methods extract some features of both videos and compare them to obtain an objective quality score. Full- and reduced-reference methods are important for the evaluation of video systems in non-real-time scenarios where on the one hand the original (reference) video data or a reduced feature data set, and on the other hand the distorted video data are available for comparison. Full-reference quality assessment methods can be used for example during the development and prototyping process of video transport systems, whereby the original video can be delivered offline to compare with the distorted video at the receiver. If only some features of the original video are available at the receiver, reduced-reference methods can be used. In contrast, no-reference methods try to assess the quality of a distorted video without any reference to the original video. These methods are commonly used for real-time quality assessments at a receiver without availability of the original video data.

Although useful, these objective quality metrics only address the perceived quality of a video session partly, since these metric do not correlate perfectly with the human perception and not all visual distortions are always taken into account.

Moreover, audible distortions and playback failures, which have an influence on the user experience, are often ignored in these objective quality metrics.

Various kinds of visual and audible distortions can be perceived by the user during video watching. Visual noise is a random variation of brightness or colour information that is not present in reality. Visual noise can be produced for example as an undesirable by-product of image capturing. Also in the audio stream, noise can be present as an undesired, random addition to the sound signal. A drop out is a small loss of data in an audio or video stream that can be caused by packet loss. Video drop outs can cause visual artefacts in a video frame; audio drop outs can produce audible distortions. A low frame rate or frame drops may be the cause of rough or choppy video. Besides, also the audio stream may be perceived as choppy or stuttering. Lip synchronization refers to the relative timing of the audio and video stream, i.e., the matching of lip movements with sung or spoken vocals. Another visual distortion is blurring. If a low-resolution video is displayed on a high-resolution screen, the video may be perceived as blurry: the video is not sharp and small details or sharp edges are difficult to distinct. Motion blur is visual as the apparent streaking of rapidly moving objects. Macro-blocking is an artefact in which objects or areas of the video image appear to be made up of small squares (macro blocks), rather than proper detail and smooth edges. Macro-blocking may be caused by problems during video transmission or by the codec during video compression. Pixelation is the effect in which individual pixels making up an image can be identified by the naked eye. This visual artefact is sometimes visible along the edges of objects. Lastly, chroma noise is an unnatural appearance in the video image due to fluctuations in colour and luminance.

Besides these audible and visual distortions, failures in the playback of the video may be noticeable for the user. E.g., problems with video playback may be visible as black screens. Besides, during playback the video may freeze, thereby not refreshing the video frames. Likewise, a limited network bandwidth may introduce rebuffering interruptions during video playback. Short loading times before the videos start playing may be important for the user. And in the case of multiple video channels, slow zapping times due to the channel change latency may be perceived as annoying.

Since objective quality metrics do not perfectly model the perceived visual quality, subjective quality measurements can be performed to assess how a video is really perceived by a viewer. These subjective quality assessments are usually conducted by asking human subjects to rate the perceived visual quality of the displayed media according to a provided quality scale [8]. Also for subjective quality assessments, different metrics exist. In the domain of communication, MOS testing (Mean Opinion Score) is predominantly used as a subjective measure of voice quality [9]. To obtain a MOS score, i.e., a numerical indication of the perceived quality, test subjects are asked to evaluate quality parameters by means of stan-

dardized scales using labels ranging from 'Excellent' to 'Bad' as defined by the ITU-T [4]. The DSIS method (Double Stimulus Impairment Scale) uses reference and test conditions, which are arranged in pairs, such that the first in the pair is the original (reference) video and the second is the same video distorted. After video watching, the test subjects are asked to vote on the second video using an impairment scale (from "impairments are imperceptible" to "impairments are very annoying") [10]. Absolute Category Rating (ACR) is a quality assessment method in which a test video is presented to the test subjects once only, without a reference. Then, the test subjects rate the quality of the videos, which should be presented in random order, on an ACR scale. ACR has been standardized in ITU-T Recommendation P.910 [11]. Also this list of subjective quality metrics is not exhaustive, but aims to denote the multiplicity of subjective measures.

Although these subjective measures better reflect the perceived quality of a service than traditional QoS parameters, they do not take into account users aspects or non-technical service aspects and as a result they are not the final goal (any more). The final goal of the optimization of a service should be to deliver a high Quality of Experience (QoE) to the user. As indicated in Figure 5.1, research regarding QoE is closely related to and partially overlapping with quality assessment methodologies, and driven by QoE indicators. The term QoE indicator refers to "a perceivable, recognized, and nameable characteristic of the individual's experience of a service which contributes to its quality" [12]. Besides these QoE indicators, the concept QoE also comprises non-technical aspects of the service. QoE considers how viewers perceive and experience multimedia content and/or multimedia communication services as a whole [13]. Since QoE relates to the user-perceived experience directly rather than to the implied impact of QoS, it is considered as a more important metric than QoS [14].

The quest for approaches that enable QoE measurement in the context of ubiquitous, 'always on' multimedia consumption is challenging but crucial. Researchers have already tried to grasp the influence of both static and more dynamic factors upon the quality of people's experiences with ICT products, applications and services for a long time. However, there is no magical formula to solve this complex problem. The increasing collaboration between researchers from different disciplines and epistemological[1] positions is in this respect not only enriching, but also necessary. In this respect, the definition of QoE is a much debated topic in the QoE community: various considerations and contributions have been made to the research domain [15–17]. Both from a theoretical and empirical perspective, this concept has been broadened over the last years. As a result, different definitions of QoE exist, but all have similar notion, referring to user satisfaction [18]. By the ITU-T, QoE is defined as "the overall acceptability of an application or service, as

[1] Epistemology is a branch of philosophy that investigates the origin, nature, methods, and limits of human knowledge.

perceived by the end-user", which might be influenced by 'user expectations' and 'context' [19].

Identifying, understanding, and quantifying the most determining aspects making or breaking the QoE of individual (or communities of) users and translating these rich insights into service and application optimization recommendations, is considered to be essential. Besides the perceived quality of a service (as can be measured by subjective quality metrics), QoE is also influenced by user aspects such as prior experiences with the service, personal skills, expectations of the user, motivation, emotions, etc [17]. Furthermore, if users are really interested in the content and are involved in the storyline of the movie, they might have a different idea about the disturbance of artefacts and artefacts might be perceived as less annoying [20].

Also the context of the user interacts with the QoE on different levels, thereby making it crucial to assess the QoE [21]. On the lowest level is the usage context: the interplay of situations in everyday life. This context level includes aspects such as the location and activity of the user, the time, the company during service usage, etc. On the highest level is the social context: the interplay between the societal structure and the action of members of society within and with the structure [17].

Last but not least, also non-technical aspects related to the service have an influence on the QoE: the price of the service, the usability of the service and the user interface, (the user's interest in) the video content that is available, enjoyment during service usage, utility of the service, and the type of device that is used.

QoE will continue to play a major role in the future development of broadcasting services and the design of multimedia applications, not the least in the dynamic mobile media domain. For video services, operators, and broadcasters, QoE has become already a service differentiator next to the number of channels or the content they offer [20]. Moreover, QoE has become a key factor in routing mechanisms and resource management schemes for network operators and IPTV providers [22].

As reflected in the importance of the QoE concept, users have become more demanding and expect that products, services, and applications address their personal and situational requirements [23], allowing them to have a good and pleasurable (quality of) experience anywhere and at any time. This is especially challenging in the mobile media domain, which is characterized by an exponential growth in the number of mobile devices, services, and applications, by the availability of various new content-delivery services and access networks, and by the massive adoption of mobile services by users. As mobile applications are used in dynamic and heterogeneous usage contexts, insights in the objective and subjective dimensions that may influence users' QoE in these contexts, have become crucial in view of QoE optimization [24].

5.3 Literature review

In this section, we provide an overview of existing work starting with studies on the influence of technical parameters on the video quality assessment. We then discuss a number of studies regarding QoE modelling in the context of streaming media services. Subsequently, the focus is on studies concerning the improvement of the QoE in the mobile application domain by approaches such as optimizing the handover process. Furthermore, this section discusses results regarding QoE research in WCDMA (Wideband Code Division Multiple Access) and UMTS (Universal Mobile Telecommunications System) networks. Finally, the pros and cons of controlled test beds as well as living lab experiments are reviewed.

5.3.1 Video quality assessment

Over the past years, numerous video quality assessment methods and metrics have been proposed with varying computational complexity and accuracy. Full-reference and reduced-reference media-layer objective video quality assessment methods are extensively classified, reviewed, and compared according to whether or not natural visual characteristics or perceptual characteristics are considered [25]. However, these metrics only measure (differences in) objective parameters of the video, which may not always be noticeable or important for the user. These metrics may be insufficient to reliably estimate end-users' subjective overall perception of the quality. As a result, these metrics do not always reflect the quality aspects that are really important for a good QoE. Therefore, the most reliable way of assessing and measuring the perceptual quality of video is conducting subjective experiments, in which human observers evaluate a series of video sequences [20].

Many of the studies on the perceptual quality of video have focused on how network-level parameters (such as delay, bandwidth, packet loss, and jitter) and video characteristics (such as codec, frame rate and resolution) affect the subjective quality of the multimedia content. Research about the effects of jitter on the perceptual quality of video indicated that jitter can degrade video quality nearly as much as packet loss [26]. Moreover, this study showed that the presence of even low amounts of jitter or packet loss results in a severe degradation in perceptual quality, while higher amounts of jitter and packet loss do not proportionally degrade the perceptual quality. The effects of present-generation video compression and communication technologies on the perceptual quality of digital video were evaluated via a subjective study by Seshadrinathan et al. [27]. This user study consisted of a large-scale subjective evaluation of video quality on a collection of videos distorted by a variety of application-relevant processes. Furthermore, the performance of several full-reference video quality assessment algorithms was evaluated and compared with the users' mean opinion scores.

In the context of streaming media services, technical parameters of the video (such as frame rate and bit rate) have been used to estimate quality perceptions and whether the service is acceptable, by using statistical modelling techniques thereby identifying the degree of influence of each technical parameter [28]. The resulting (classification) model predicts the user's perception of the service quality, which is considered as a QoE indicator, based on the technical characteristics of the video. This model does not take into account network-level parameters (such as packet loss, delay, jitter) but allows network operators to anticipate the user's experience and then allocate network resources accordingly [28].

For communication and entertainment systems that involve streaming media, such as teleconference applications, UDP (User Datagram Protocol) based streaming protocols such as RTP (Real-time Transport Protocol) are commonly used. These streaming media require timely delivery of information and allow no retransmissions in case of packet loss, which may lead to noticeable distortions for the user. For UDP based streaming protocols, QoE is determined as a function of the technical video parameters (resolution, frame rate and codec) and spatial and temporal video artefacts resulting from network imperfections, e.g., packet loss, delay, and jitter [29]. Although intelligent mechanisms in core and distribution networks may prevent congestion and packet loss, video streaming over IP (Internet Protocol) networks is error-prone and subject to a wide range of distortions, artefacts, and degradations during transmission [30].

The perceptual effects of packet loss during UDP based streaming are depending on various technical parameters of the video, such as the type of frame in which packet loss occurs (I, P, or B-frame), the resolution, and the video codec as well as on network characteristics such as burst patterns. For MPEG encoded videos for example, packet-loss rates as low as 3% might induce much higher frame error rates (up to 30% of frame errors) and clearly-noticeable distortions for video with an approximate duration of five mintues [31]. Compared to a perfect video, perceptual quality drops by over 50% in the presence of a packet-loss rate of about 8%. Additional packet loss, up to 22%, induces a further (but weaker) decline in the perceptual quality [26]. Whereas these studies discussed the influence of packet loss on the perceptual quality in the context of video watching on desktop computers or television sets (i.e., a fixed setting), Chapter 6 of this dissertation investigates this influence in a mobile context, which can have a serious impact on the QoE.

As an alternative for RTP, nowadays a lot of video content is available via Dynamic Adaptive Streaming over HTTP (DASH), a technique also known as progressive download which is based on HTTP (Hypertext Transfer Protocol) and TCP (Transmission Control Protocol) which assures a reliable, ordered delivery of video packets. Using progressive download, buffering mechanisms and packet retransmissions can avoid the audiovisual distortions due to packet loss and jitter,

but may incur rebuffering interruptions and additional start-up delays compared to UDP based streaming applications [32]. In other words, in case of a network bottleneck the displayed content does not suffer from video quality degradation, but playback suffers from rebuffering interruptions.

The impact of the underlying transport protocol on the QoE for streaming media services was studied by comparing UDP based and TCP based video streaming [33]. The results indicated that TCP based video-on-demand delivery, which is for instance used by YouTube, outperforms UDP based video streaming in terms of the user's perceived quality for network bottleneck scenarios. The user's personal experience rating showed to be highly influenced by the number of video interruptions during video transmission over TCP. All users rated their video experience with the maximum rating if no interruption occurred. In contrast, in case of two and more rebuffering interruptions, more than 30% of the users rated the video experience with the lowest rating score. From these results, the authors concluded that users tolerate one interruption of 3 seconds per clip but more interruptions, especially more than two, significantly reduce the user's perceived video quality and experience. Since user expectations and experiences might be different for mobile applications due to other characteristics of the hardware (e.g., type of device or screen size) and the wireless data networks (e.g., cellular networks), we assume that these conclusions do not apply (without adjustments) to the acceptability regarding video interruptions on the mobile platform, which is investigated in our research.

5.3.2 QoE on the mobile platform

The proliferation of mobile multimedia applications over wireless, resource-constrained networks has raised the need for methods that adapt these applications both to network resource constraints and to clients' QoE requirements [34]. Various studies have been devoted to the measurement of QoE by considering both measurable (e.g., waiting time) and non-measurable (e.g., the user's expectations) parameters in the domain of mobile multimedia services. Using these parameters, it is possible to produce quantifiable quality metrics for QoE evaluation of multimedia streaming by performing analysis at the application and network levels [35]. The so-called 'non-measurable' parameters are related to the users' evaluation of the investigated QoE features, their behaviour, expectations, emotions, etc. Increasingly, interdisciplinary research is set up in order to take these human aspects into account.

Several researchers have studied the QoE of mobile media sessions in cellular data networks. Based on experiments in WCDMA networks, a predictive QoE model for multimedia applications was proposed [36]. A stepwise regression analysis revealed the most relevant factors for the QoE: the number of transmission errors, buffering occurrences, and coding profile. Moreover the study pointed

to the importance of the buffering duration and frequency. Experiments in UMTS networks found that the effect of the RTT (round-trip time) and bandwidth are very perceivable by the users while browsing web pages [37]. The same study showed that the initial start-up time of streaming video, which is influenced by these network parameters, is crucial, independent of the quality of the streaming. The test subjects were also very sensitive to any rebuffering that occurs after the streaming has started, and rated the overall quality regardless of the video quality after the rebuffering interruption.

Other studies considered the rebuffering length and rebuffering frequency as the properties that have the greatest impact on QoE. If interruption is unavoidable, a single rebuffering is a better solution than repeated rebuffering events [38]. Other subjective tests showed that also a single rebuffering interruption can reduce the users' QoE considerably [39]. A very recent study compared the impact of initial delays vs. interruptions and found that the latter should always be avoided, even at the cost of increasing the initial waiting time due to prebuffering [40]. Regrettably, these studies do not evaluate how much time can be spent on the rebuffering of mobile video before this becomes unacceptable for the user, or in other words the acceptability regarding the initial loading time and the rebuffering interruptions during video playback. Therefore, Chapter 6 of this research investigates the acceptability of rebuffering interruptions during mobile video watching and provides a model that estimates this acceptability considering the initial loading time and rebuffering interruptions.

In the context of heterogeneous mobility[2], it is also investigated how network hand-overs can be optimized and made seamless, allowing the user to have the best possible experience. To improve this handover process across multiple link-layer access technologies, a modified Android user terminal using the IEEE 802.21 framework has been proposed [41]. The assessment of the handover process via an experimental test bed showed that under the proposed solution the handover delay and packet loss are significantly lower than the ones resulting from the normal operation.

5.3.3 Controlled lab environment vs. living lab

To evaluate QoE in the context of mobile applications or services, both traditional test beds with controlled parameters or living lab experiments in the field can be set up.

Although living lab experiments are an extension towards more natural and realistic research test environments [42], a strong tradition exists in experimental research taking place in test beds with controlled parameters [43, 44]. The controlled laboratory settings allow for transparent, rigorous, and replicable testing

[2]heterogeneous mobility allows the movement of mobile devices to other networks in order to fulfil the service requirements

of new technologies, scientific theories, and tools regarding the quantification and optimization of the QoE. Research using this kind of test beds makes it possible to investigate the relative influence of particular isolated parameters on users' quality perceptions.

The experiment presented in Chapter 6 of this research has been carried out in a controlled environment test room. The experiment investigates if and how users' QoE is influenced by the number of rebuffering interruptions during mobile video watching.

Yet, especially when the focus is on 'ubiquitous QoE' and its interplay with dynamic contextual and user-related variables, the complementary value of more ecologically valid approaches should be explored. (The ecological validity of a study denotes the extent to which the methods, materials, and setting of the study approximate the real-world that is being examined [45]). Although no common or standardized methodologies have been developed in this respect, interesting work has already been done in this area, e.g., in the domain of pervasive computing [46], and mobile TV [47]. Various researchers pointed to the relevance of the living labs approach for "integrating technology components into the complex environment of the wireless world and end-users in their daily life" [48]. In contrast to controlled laboratory settings, living lab experiments are less transparent and predefined but aim to provide more natural settings for studying QoE by involving the users in the innovation process [49].

In the definition of Følstad [42], living labs are "environments for innovation and development where users are exposed to new ICT solutions in (semi-)realistic contexts, as part of medium- or long-term studies targeting evaluation of new ICT solutions and discovery of innovation opportunities". Drawing on the open and user-driven innovation rationale, the living lab approach might help to facilitate the continuous and systematic involvement of end-users and to enable researchers to understand the drivers and barriers of QoE in heterogeneous real-life contexts [49]. Moreover, as living labs 'bring the lab to the people' and draw on 'real' experiences from 'real' users, QoE research in such settings will likely yield more trustworthy results and have a higher ecological validity than research in controlled environments [50].

In this respect, Staelens et al. compared QoE assessment performed in controlled laboratory environments and in the natural setting of people's everyday life context [20]. They discovered significant differences concerning perceived artefacts and acceptability of the video quality. In general, artefacts showed to be less perceptible during real-life QoE assessment. So, conclusions which are obtained using a standardized subjective-quality assessment methodology may not always hold on the case of real-life QoE assessment since user expectations and context influence QoE.

The experiments presented in Chapter 7 and 8 of this research have been carried out in a living lab. Chapter 7 presents a model for the subjective evaluations of various quality aspects during mobile video watching, such as the perceived distortion and loading speed. Chapter 8 discusses the influence of QoE on the user's explicit feedback that is used for generating recommendations. In this way, this chapter links the research of the first part of this dissertation concerning recommender systems, and the second part, that is about QoE.

5.4 Conclusion

Entertainment and multimedia are the key functionalities in emerging mobile markets. The ability to understand and quantify the QoE, will play a major role in the success of these mobile services. Over the past years, a number of studies have looked into how distortions or interruptions influence QoE. However, systematic research investigating the users' evaluation and acceptability with respect to distortions, loading time, and rebuffering interruptions during mobile video watching is still rather limited. Moreover, results based on fixed video watching (using wired devices) cannot be applied to the mobile domain without adjustments because the user's expectations can differ depending on the platform, and because the user's experience is influenced by the type of device and display. This dissertation contributes to the research area of QoE by quantifying the perceived distortion and loading speed, and the users' acceptability thresholds with respect to rebuffering interruptions in a mobile environment.

References

[1] J. O'Halloran. *Live and downloaded mobile video grow strongly*. Technical report, RapidTVNews, 2011. Online available at `http://www.rapidtvnews.com/index.php/2011033011228/video-increasingly-crucial-to-mobile-development.html`.

[2] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011–2016*. Technical report, Cisco Inc., 2012. Online available at `http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html`.

[3] K. Mun. *Mobile video calling: improving QoE. Techzine, Technology And Research E-Zine*. Technical report, Alcatel-Lucent, 2011. Online available at `http://www2.alcatel-lucent.com/blogs/techzine/2011/mobile-video-calling-improving-qoe/`.

[4] *E.800: Terms and definitions related to quality of service and network performance including dependability*. Technical report, ITU-T Recommendation, International Telecommunication Union, 1994. Updated September 2008 as Definitions of terms related to quality of service.

[5] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam. *Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison*. IEEE Transactions on Broadcasting, 57(2):165–182, June 2011.

[6] M. Vranjes, S. Rimac-Drlje, and D. Zagar. *Objective video quality metrics*. In ELMAR, 2007, pages 45 –49, September 2007.

[7] J. Valerdi, A. Gonzalez, and F. Garrido. *Automatic Testing and Measurement of QoE in IPTV Using Image and Video Comparison*. In Fourth International Conference on Digital Telecommunications, 2009. ICDT '09, pages 75 –81, July 2009.

[8] L. Karam, T. Ebrahimi, S. Hemami, T. Pappas, R. Safranek, Z. Wang, and A. Watson. *Introduction to the Issue on Visual Media Quality Assessment*. IEEE Journal of Selected Topics in Signal Processing, 3(2):189–192, April 2009.

[9] T. de Koning, P. Veldhoven, H. Knoche, and R. Kooij. *Of MOS and men: bridging the gap between objective and subjective quality measurements in mobile TV*. In Multimedia on Mobile Devices, February 2007.

[10] S. Grgić, M. Grgić, and M. Mrak. *Reliability of objective picture quality measures*. Journal of Electrical Engineering, 55:3–10, 2004.

[11] *P.910 : Subjective video quality assessment methods for multimedia applications*. Technical report, ITU-T, International Telecommunication Union, 2008.

[12] P. Le Callet, S. Möller, and A. Perkis (eds.). *Qualinet White Paper on Definitions of Quality of Experience*. Technical report, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, June 2012. Version 1.1.

[13] U. Reiter. *Overall perceived audiovisual quality - What people pay attention to*. In IEEE 15th International Symposium on Consumer Electronics 2011 (ISCE), pages 513–517, June 2011.

[14] L. A. Rowe and R. Jain. *ACM SIGMM retreat report on future directions in multimedia research*. ACM Transactions on Multimedia Computing, Communications, and Applications, 1(1):3–13, 2005.

[15] P. Reichl. *From charging for Quality of Service to charging for Quality of Experience*. Annals of telecommunications - Annales des télécommunications, 65:189–199, 2010.

[16] S. Moller, K.-P. Engelbrecht, C. Kuhnel, I. Wechsung, and B. Weiss. *A taxonomy of quality of service and Quality of Experience of multimodal human-machine interaction*. In International Workshop on Quality of Multimedia Experience, QoMEX 2009, pages 7–12, July 2009.

[17] D. Geerts, K. De Moor, I. Ketykó, A. Jacobs, J. Van den Bergh, W. Joseph, L. Martens, and L. De Marez. *Linking an integrated framework with appropriate methods for measuring QoE*. In Second International Workshop on Quality of Multimedia Experience, QoMEX 2010, pages 158 –163, June 2010.

[18] D. Soldani, M. Li, and R. Cuny. *QoS and QoE Management in UMTS Cellular Systems*, pages i–xxvii. John Wiley & Sons, Ltd, 2006.

[19] *Definition of Quality of Experience (QoE)*. Liaison statement, ITU-T, International Telecommunication Union, 2007. Ref.: TD 109 rev 2 (PLEN/12).

[20] N. Staelens, S. Moens, W. Van den Broeck, I. Mariën, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester. *Assessing Quality of Experience of IPTV and Video on Demand Services in Real-Life Environments*. IEEE Transactions on Broadcasting, 56(4):458–466, December 2010.

[21] G. Mantovani. *Social Context in HCI: A New Framework for Mental Models, Cooperation, and Communication*. Cognitive Science, 20(2):237–269, 1996.

[22] S. Balasubramaniam, J. Mineraud, P. McDonagh, P. Perry, L. Murphy, W. Donnelly, and D. Botvich. *An Evaluation of Parameterized Gradient Based Routing With QoE Monitoring for Multiple IPTV Providers*. IEEE Transactions on Broadcasting, 57(2):183 –194, June 2011.

[23] T. De Pessemier, K. De Moor, I. Ketykó, W. Joseph, L. De Marez, and L. Martens. *Investigating the influence of QoS on personal evaluation behaviour in a mobile context*. Multimedia Tools and Applications, 57:335–358, 2012.

[24] L. De Marez and K. De Moor. *The Challenge of User- and QoE-Centric Research and Product Development in Today's ICT-Environment*. In The Good, The Bad and The Unexpected. The user and the future of innovation and communication technologies, volume 1. The Icfai University Press, 2007.

[25] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam. *Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison*. IEEE Transactions on Broadcasting, 57(2):165 –182, June 2011.

[26] M. Claypool and J. Tanner. *The effects of jitter on the perceptual quality of video*. In Proceedings of the seventh ACM international conference on Multimedia (Part 2), MULTIMEDIA '99, pages 115–118, New York, NY, USA, 1999. ACM.

[27] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack. *Study of Subjective and Objective Quality Assessment of Video*. IEEE Transactions on Image Processing, 19(6):1427–1441, June 2010.

[28] F. Agboma and A. Liotta. *QoE-aware QoS management*. In Proceedings of the 6th International Conference on Advances in Mobile Computing and Multimedia, MoMM '08, pages 111–116, New York, NY, USA, 2008. ACM.

[29] K. Piamrat, C. Viho, J.-M. Bonnin, and A. Ksentini. *Quality of Experience Measurements for Video Streaming over Wireless Networks*. In Sixth International Conference on Information Technology: New Generations, 2009. ITNG '09, pages 1184 –1189, April 2009.

[30] J. Asghar, F. Le Faucheur, and I. Hood. *Preserving Video Quality in IPTV Networks*. IEEE Transactions on Broadcasting, 55(2):386 –395, June 2009.

[31] J. M. Boyce and R. D. Gaglianello. *Packet loss effects on MPEG video sent over the public Internet*. In Proceedings of the sixth ACM international conference on Multimedia, MULTIMEDIA '98, pages 181–190, New York, NY, USA, 1998. ACM.

[32] R. Mok, E. Chan, and R. Chang. *Measuring the quality of experience of HTTP video streaming*. In 2011 IFIP/IEEE International Symposium on Integrated Network Management (IM), pages 485–492, May 2011.

[33] T. Hoßfeld, R. Schatz, T. Zinner, M. Seufert, and P. Tran-Gia. *Transport Protocol Influences on YouTube Videostreaming QoE*. Science research report series, University of Würzburg Institute of Computer, 2011. Online available at `http://www3.informatik.uni-wuerzburg.de/TR/tr482.pdf`.

[34] W. He, K. Nahrstedt, and X. Liu. *End-to-end delay control of multimedia applications over multihop wireless links*. ACM Transactions on Multimedia Computing, Communications, and Applications, 5(2):16:1–16:20, 2008.

[35] A. Perkis, S. Munkeby, and O. Hillestad. *A model for measuring Quality of Experience*. In Proceedings of the 7th Nordic Signal Processing Symposium, 2006. NORSIG 2006, pages 198–201, June 2006.

[36] O. Bradeanu, D. Munteanu, I. Rincu, and F. Geanta. *Mobile Multimedia End-User Quality of Experience Modeling*. In International Conference on Digital Telecommunications, 2006. ICDT '06, page 49, August 2006.

[37] O. Teyeb, T. B. Sørensen, P. Mogensen, and J. Wigard. *Subjective evaluation of packet service performance in UMTS and heterogeneous networks*. In Proceedings of the 2nd ACM international workshop on Quality of service & security for wireless and mobile networks, Q2SWinet '06, pages 95–102, New York, NY, USA, 2006. ACM.

[38] X. Tan, J. Gustafsson, and H. Gunnar. *Perceived video streaming quality under initial buffering and rebuffering degradations*. In Proceedings of the MESAQIN conference, June 2006.

[39] J. Gustafsson, G. Heikkila, and M. Pettersson. *Measuring multimedia quality in mobile networks with an objective parametric model*. In 15th IEEE International Conference on Image Processing, 2008. ICIP 2008, pages 405–408, October 2008.

[40] T. Hoßfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen. *Initial delay vs. interruptions: Between the devil and the deep blue sea*. In Fourth International Workshop on Quality of Multimedia Experience, QoMEX 2012, pages 1–6, July 2012.

[41] R. Silva, P. Carvalho, P. Sousa, and P. Neves. *Enabling Heterogeneous Mobility in Android Devices*. Mobile Networks and Applications, 16:518–528, 2011.

[42] A. Følstad. *Living Labs for Innovation and Development of Information and Communication Technology: A Literature Review*. Electronic Journal of Organizational Virtualness, 10:99–131, 2008.

[43] *P.911: Subjective audiovisual quality assessment methods for multimedia applications*. Technical report, ITU-T, International Telecommunication Union, 1998. Online available at `http://www.itu.int/rec/T-REC-P.911-199812-I/en`.

[44] *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment*. Technical report, ITU-T, International Telecommunication Union, 2003. Online available at `http://www.itu.int/md/T01-SG09-C-0060/en`.

[45] M. B. Brewer. *Research design and issues of validity.* In H. T. Reis and C. M. Judd, editors, Handbook of research methods in social and personality psychology, pages 3–16. Cambridge University Press, New York, NY, USA, 2000.

[46] L. Li-yuan, Z. Wen-an, and S. Jun-de. *The Research of Quality of Experience Evaluation Method in Pervasive Computing Environment*. In 1st International Symposium on Pervasive Computing and Applications, 2006, pages 178–182, August 2006.

[47] S. Jumisko-Pyykkö and M. M. Hannuksela. *Does context matter in quality evaluation of mobile television?* In Proceedings of the 10th international conference on Human computer interaction with mobile devices and services, MobileHCI '08, pages 63–72, New York, NY, USA, 2008. ACM.

[48] M. Ponce de Leon, M. Eriksson, S. Balasubramaniam, and W. Donnelly. *Creating a distributed mobile networking testbed environment-through the Living Labs approach*. In 2nd International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities, 2006. TRIDENTCOM 2006, pages 135–139, 2006.

[49] K. De Moor, I. Ketykó, W. Joseph, T. Deryckere, L. De Marez, L. Martens, and G. Verleye. *Proposed Framework for Evaluating Quality of Experience in a Mobile, Testbed-oriented Living Lab Setting*. Mobile Networks and Applications, 15:378–391, 2010.

[50] J. Schumacher and V.-P. Niitamo (eds.). *European living labs - a new approach for human centric regional innovation.* Wissenschaftlicher Verlag Berlin, 2008.

6

QoE research in a controlled laboratory environment

6.1 Introduction

This chapter describes the results of a controlled laboratory experiment that explores the thresholds at which the technical quality of a mobile video service becomes unacceptable for users. A subjective experiment drawing on the logging of technical parameters combined with subjective evaluations by a user panel resulted in a model for quantifying the acceptability of video interruptions. The results of this analysis provide insights into the QoE and (un)acceptability regarding video interruptions for different network conditions and video parameters. The conclusions of this chapter can be used as a guideline for service design and network dimensioning.

6.2 Test setup

6.2.1 Goals of the study

The main aim of this study is to investigate the influence of rebuffering interruptions on QoE during mobile video watching. More concretely, we investigate if and how the test subjects' QoE is influenced by the number of rebuffering interruptions in six technical scenarios combining three simulated connection types (low, medium, and high bandwidth) and two video qualities (low and high quality).

We investigate the influence of the objective measures mentioned in Table 6.3 on different measures of QoE, as dependent variables in our study. These include the overall experience rating, and the evaluation of both the overall technical quality, as well as specific QoE indicators (see Chapter 5), being interruptions, loading time, and fluidity (sometimes also referred to as fluentness or smoothness of the video playback). We included these specific QoE indicators to investigate their relative importance and thorough evaluation by test subjects and because previous research pointed to their importance. Finally, we investigate possible differences in terms of the acceptability of video playback interruptions due to rebufferings. Following the definition given in [1], acceptability refers to "a binary measure to locate the threshold of minimum quality that fulfils user quality expectations and needs for a certain application or system". In addition, we also take into account the importance of specific QoE indicators related to mobile video watching before and after the actual test. Furthermore, we complement the test subjects' ratings with qualitative feedback on expectations and importance of features and influencing factors, indicated by test subjects themselves. We now describe the technical and experimental setup in more detail.

6.2.2 Procedure

In this section, we provide the details about the experiment, which consisted of three successive phases:

6.2.2.1 Phase 1: pre- questionnaire & instruction meetings

Before the actual experiment started, participants were asked to fill in a traditional paper questionnaire consisting of closed and open questions. The questionnaire inquired after their socio-demographic characteristics, type and connection possibilities of their current mobile phone, and experiences and habits (in terms of viewing frequency, ranging from never to several times a day) regarding the watching of video content on a mobile phone. Next, by means of the first open question, the test subjects with prior experience were asked to specify which characteristics (QoE indicators) related to mobile video watching they personally consider to be essential for having a good experience. 'Open' in this context means that no pre-defined answer categories were given and that test subjects were able to express themselves in their own words. Thereupon, the test subjects who had no prior experience with mobile video watching were able to indicate what according to them might influence their experience. By means of this second open question, we wanted to gain more insight in the test subjects' expectations with regard to possible influencing factors. Finally, the test subjects were asked to indicate how (un)important they considered a number of listed aspects in order to have a good experience during the watching of video content on a mobile phone. These aspects

and their importance for mobile video sessions, assessed on a 5-point rating scale going from 1 (not important at all) to 5 (very important), are listed in Table 6.5.

After this preliminary questionnaire, test subjects received instructions on how to switch on/off the device (Google Nexus One running on Android 2.1), how to use the touch screen, how to access the test application, and how to select and watch the videos. Since each video watching is followed by a small questionnaire on the device, test subjects were also shown how to fill in this electronic questionnaire using the touch screen and given instructions concerning the interpretation of the questions and operational definitions of the QoE measures. After this briefing session, every test subject was given a device and asked to watch 14 videos, each with a length of approximately two minutes, in a controlled environment (i.e., the research lab of our university).

6.2.2.2 Phase 2: mobile video watching in a controlled laboratory environment

Figure 6.1 shows the architecture of the video delivery system used in the controlled laboratory experiment, consisting of the client device (i.e., a smartphone running the video player), the video server offering the content, and the technical database storing the objective parameters and subjective evaluations.

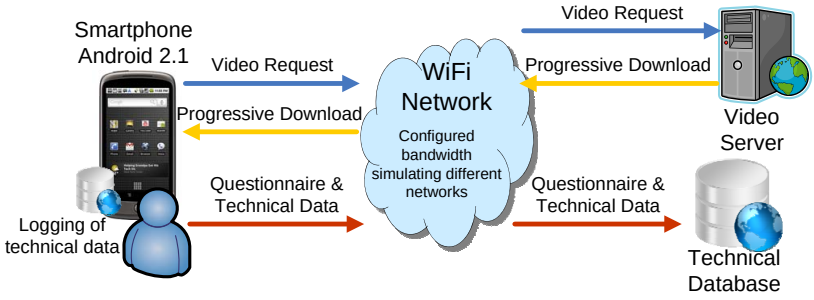


Figure 6.1: The architecture of the video delivery system used in the controlled laboratory experiment

During the setup of the experiment, the 14 videos were preselected from a large content pool and hence cover a large variety of genres including entertainment, technology, music, film, animation, science, cartoons, and news. Since progressive download is used as transport protocol in this experiment, video playback can start before the video file is completely downloaded to the device. For this video playback, the standard video player of the Android operating system is used. The videos are transmitted to the device over a WiFi connection (802.11g) of which the maximum available bandwidth per device can be configured. By limiting the band-

width of the WiFi connection, (the bandwidth of) different cellular data networks can be simulated.

Table 6.1 lists the three different connection types used in this experiment, together with their theoretical and measured throughput (i.e., the average rate of successful data delivery over the communication channel) and the standard deviation of the measured throughput. The theoretical throughput equals the maximum available bandwidth per device as configured in the wireless access point during the experiment. Because of protocol overhead, the actually measured throughput is slightly lower than the theoretical throughput. This actual throughput is calculated by averaging ten measurements of the download speed, which are performed by an application called SpeedTest [2] before the start of the application. The standard deviation is calculated based on these ten measurements of the actual throughput. Since the download speed is far more important than the upload speed for video reception on the device, only the download speed was considered. The low standard deviation of the measurements indicates that the throughput was only slightly varying during the experiment.

Connection type	Theoretical maximum throughput (kbit/s)	Average of the measured throughput (kbit/s)	Standard deviation of the measured throughput (kbit/s)
High-bandwidth	54000	14853	342
Low-bandwidth	750	564	9
Medium-bandwidth	1800	1516	44

Table 6.1: Theoretical and measured throughput of the different connection types

The high-bandwidth connection used in this experiment represents a WiFi communication channel and has no imposed restrictions. Although the device does not use the full theoretical bandwidth of the wireless connection, the measured throughput is representative for a WiFi connection and sufficient for all mobile services. The low-bandwidth connection of this experiment has a throughput that is typical for a UMTS connection since the planned transceiver capacity of a UMTS network is typically ranging from 400 kbits/s to 700 kbits/s [3]. At the time of the study, Belgian network operators planned HSPA (High Speed Packet Access) networks to provide each mobile user with a bandwidth capacity of 1.5 Mbit/s in the downlink channel (this information is based on confidential interviews with an operator). So, the medium-bandwidth connection of the experiment has a throughput that is typical for an HSPA connection that is available for end-users in Belgium. Although real mobile data networks (such as UMTS or HSPA networks) and the proposed shaped WiFi network have a different behaviour regarding packet loss and jitter, these differences are hidden by packet retransmis-

sions and data buffering of the reliable transport protocol (TCP/HTTP) that was used in this experiment.

These three connection types enable us to investigate the influence of network throughput on the subjectively evaluated experience, technical quality, and acceptability of a video session. To demonstrate the reference quality of the mobile videos, two videos were transmitted to the mobile device using a high-bandwidth connection. Since this high-bandwidth connection has no network limitations influencing the audiovisual quality during video playback, the quality of the video source is the only variable that affects the quality of the video playback. Therefore, this case is less interesting to study and so the number of videos using this high-bandwidth connection is limited to two in the experiment. The other 12 videos that test subjects had to watch were transmitted over a low- or medium-bandwidth connection, which may introduce video interruptions during playback (6 videos for each connection type).

To investigate the influence of the quality of the video source on the user's QoE during video watching, videos are transcoded into two different quality versions. Table 6.2 lists the characteristics of these two video versions and shows that the high-quality version has a higher resolution, bit rate, and frame rate compared to the low-quality version. During playback, both quality versions were upsampled by the device and displayed in full-screen. For both versions, the ITU-standard (International Telecommunication Union) H.264 AVC (advanced video coding) is used, since it is currently one of the most commonly used formats for the recording, compression, and distribution of (high definition) video [4]. The AAC LC 3 (Advanced Audio Coding, Low Complexity profile 3) compression scheme is used for the audio track. The average audio bit rate of 62kbit/s is rather low, but satisfactory for streaming video on the mobile devices given the moderate quality of the speakers of the smartphone. No noticeable disturbances were audible in the sound. Since test subjects did not have to evaluate the audio quality separately in the experiment, all videos are coded with the same audio bit rate. For each connection type, as many low-quality videos as high-quality videos are used in the experiment. To avoid boredom, the test subjects had to watch all videos only once during the experiment. So summarized, test subjects had to watch 2 videos without network limitations (1 in low quality, 1 in high quality), 6 videos transmitted using a medium-bandwidth connection (3 in low quality, 3 in high quality) and 6 videos transmitted using a low-bandwidth connection (3 in low quality, 3 in high quality).

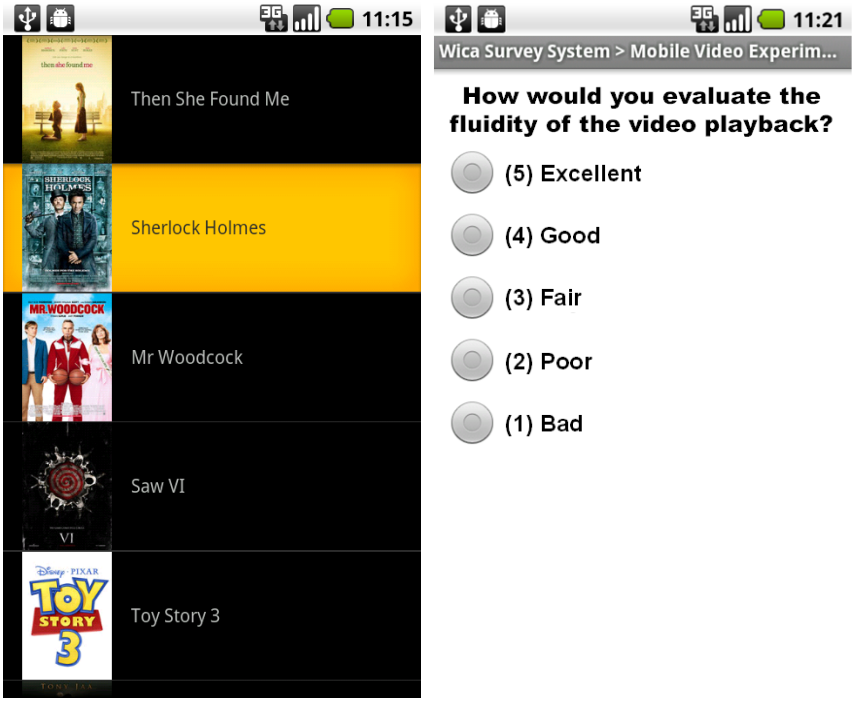
The test subjects were not informed about these changing network characteristics and the variable quality of the video source but received a list of videos with just a thumbnail and the title as additional information (Figure 6.2(a)). Selecting a video from this list starts the transmission to the mobile device and the playback of that video. The videos were selected and watched by the test subjects in the order

Low Quality Video Source			
Audio		Video	
Codec	AAC LC 3	Codec	H.264/AVC
Average bit rate	62 kbit/s	Average bit rate	109 kbit/s
Maximum bit rate	81 kbit/s	Maximum bit rate	507 kbit/s
Channels	2	Resolution	256 x 144
Sampling frequency	44100 Hz	Frame rate	13 fps
High Quality Video Source			
Audio		Video	
Codec	AAC LC 3	Codec	H.264/AVC
Average bit rate	62 kbit/s	Average bit rate	765 kbit/s
Maximum bit rate	81 kbit/s	Maximum bit rate	1815 kbit/s
Channels	2	Resolution	512 x 288
Sampling frequency	44100 Hz	Frame rate	25 fps

Table 6.2: Technical parameters of the mobile video used in the controlled laboratory environment

they prefer, at a fixed location in the laboratory. Each test subject received the same list of videos and each of these videos had a predefined quality and transmission condition which remained the same in every test. The videos of each connection type / quality combination are covering a variety of content genres to ensure that there is no link between on the one hand the content and on the other hand the quality of the video source or the bandwidth of the communication channel.

During each video playback, various technical parameters regarding the network and video are logged. Table 6.3 shows these measured objective parameters with their unit, value, and sampling rate. For each video, the bandwidth of the communication channel and the quality of the video source were determined during the setup of the experiment. The loading time, which is also measured for each video playback, is defined as the time between selecting a video and the moment when the video starts playing. During the playback of the video, multiple rebufferings may be required. The rebuffering time is defined as the time period that video playback is interrupted because the video buffer is (almost) empty and waiting for new data from the network connection. The loading time and rebuffering time are used to investigate the subjective acceptability of video playback interruptions (Section 6.3.4). Through an application called Wireshark [5], the mean RTT (round-trip delay time) is measured during each video playback. Wireshark defines the RTT as "the difference in capture time of TCP packets with a certain sequence number and the corresponding follow-up acknowledgement packets from



(a) The video selection mechanism

(b) The questionnaire that is shown after each video

Figure 6.2: Screenshots of the video application on the mobile device

the receiver". The measured RTT, enables us to investigate the influence of the network delay on the subjective user evaluations (Section 6.3.3).

Parameter	Unit	Value	Sampling Rate
Video ID	Integer	[0, ∞[	Each video
Video quality	State	{low, high}	Each video
Bandwidth	State	{low, medium, high}	Each video
Loading time	Seconds	[0, ∞[	Each video
Rebuffering times	Seconds	[0, ∞[	Each rebuffering
Mean RTT	Seconds	[0, ∞[	Each video

Table 6.3: The measured objective parameters of the video sessions in the controlled laboratory experiment

After each video playback, a short questionnaire pops up on the screen of the mobile device. Figure 6.2(b) shows a screenshot of this digital questionnaire that

test subjects were asked to fill in immediately after watching a video. Through this feedback form, test subjects can evaluate the content of the videos, their general experience, the general technical quality and specific features, and finally, the acceptability of the video quality. After the evaluation of the content itself, test subjects were firstly asked to rate the overall technical quality of the video. In the briefing for the test subjects, the following operational definition was given: "By technical quality, we mean the overall quality of the different technical features that you - as a viewer - can perceive (these include, e.g., the sharpness of the image, the synchronization between the sound and image, the fluidity of the video, loading speed, visual artefacts or errors in the video, ...). Other aspects, such as the appreciation of the content of the video, are not part of this technical quality". A high score corresponds with a positive evaluation of the technical quality; a low score indicates that the user is not at all or not really satisfied with the technical quality. Then, separate questions were provided to assess a number of specific QoE indicators, being the impact of interruptions, loading speed, and fluidity. Interruptions were explained as undesired pauses or breaks in the video playback. The loading speed is evaluating the waiting time between selecting a video and the start of the video playback. Fluidity was explained to the test subjects as the degree to which the images follow up on each other without delay, interruptions or freezes.

The choice of the rating scale might be seen as an important element in the subjective testing methodology. Nevertheless, a direct comparison between four different rating scales based on experimental data showed no overall statistical differences between the different scales [6]. Table 6.4 lists the questions and the used measurement scales as recommended by ITU-T [7], i.e., a 5-level subjective quality evaluation scale (1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent). For the question regarding video interruptions, the labels of the evaluation scale were as follows: 1 = Noticeable, very annoying, 2 = Noticeable, annoying, 3 = Noticeable, slightly annoying, 4 = Noticeable, not annoying, 5 = Not noticeable. Both the numbers of the measurement scale and the corresponding labels were shown to the test subjects.

A limitation of the followed approach is linked to the instructions given to subjects to focus on specific QoE indicators. This may have biased the obtained results to some degree as test subjects may have become more aware of and sensitive to these aspects. However, this possible bias is inherent to this type of setup and hard to avoid. Moreover, without clear instructions and tasks, the results may also be biased as test subjects might be basing their evaluation and ratings on completely different aspects, or have a different understanding of specific QoE indicators.

Reference	Question	Possible answers
1. Content	How would you evaluate the content of the video?	5-point rating scale: 1 = Bad; 5 = Excellent
2. Technical Quality	How would you evaluate the technical quality of the video in general?	5-point rating scale: 1 = Bad; 5 = Excellent
3. Interruptions	Did you experience distortions or interruptions as annoying during video playback?	5-point rating scale: 1 = Noticeable, very annoying; 5 = Not noticeable
4. Fluidity	How would you evaluate the fluidity of the video playback?	5-point rating scale: 1 = Bad; 5 = Excellent
5. Loading	How would you evaluate the loading speed of the video?	5-point rating scale: 1 = Bad; 5 = Excellent
6. Experience	How would you evaluate your general experience during video playback?	5-point rating scale: 1 = Bad; 5 = Excellent
7. Acceptability	Would you evaluate the technical quality of this video as acceptable?	Binary answer: a) Acceptable, b) Not acceptable

Table 6.4: The questions that were used to evaluate the video immediately after the playback in the controlled laboratory experiment, together with a reference to these questions and the possible answers

6.2.2.3 Phase 3: post-questionnaire

After the subjective experiment, test subjects were asked again to evaluate the importance of the aspects of Table 6.5 with respect to a good experience during mobile video watching. Given the variable quality of the video source and the variable bandwidth of the network connection, test subjects might have changed their opinion about the importance of the various technical aspects of mobile video. Additionally, there was an open question where the test subjects could indicate the three aspects that according to them are most essential in view of having a good QoE in the context of mobile video watching.

6.2.3 Sample description

In total, 12 sessions were organized (in groups of maximum five test subjects), since five Nexus One devices, running on Android 2.1 as operating system, rotated among the test subjects. During a period of two weeks, 57 test subjects (38 men and 19 women), selected using a convenience panel-sampling method, participated in the experiment. The mean age of the participants is 29.5 with a standard

deviation of 5.2. As most of them work or study at the university, the sample is composed of researchers, project managers, students, secretaries, and maintenance personnel. Notwithstanding the technical background of some test subjects, the question about their habits regarding mobile video watching showed that many of them had no prior experiences with mobile video (Section 6.3.1). As a result, we believe that the influence of the test subjects' background on their subjective evaluations during the experiment is rather limited.

After checking the data in terms of their completeness, the technical data and subjective evaluations from the questionnaire were coupled and integrated into one data file, containing 785 samples, which could be used for further analysis and which is enough for drawing statistically-founded conclusions [8].

6.3 Results

We first discuss the results of the pre- and post-questionnaire in Section 6.3.1. Thereupon, Section 6.3.2 investigates the differences in terms of the objective measures for each combination of connection type and source quality of the video. Section 6.3.3 discusses the differences in terms of the subjective measures and the correlation between the objective and subjective measures. Section 6.3.4 elaborates further on these subjective measures and investigates which combinations of connection type and source quality receive a significantly different evaluation regarding technical quality and QoE. Finally, Section 6.3.5 discusses the acceptability of the technical quality and the influence of rebufferings on this acceptability.

6.3.1 Pre- and post-questionnaire

Figure 6.3 shows a pie chart visualizing the types of mobile phone, characterized by their technical capabilities, and the number of test subjects owning such a device: the majority of the test subjects (31 of the 57) owns a smartphone (with or without touch screen) enabling them to watch mobile video. However, a question regarding mobile video consumption indicated that many of these smartphone users never use their phone for watching mobile video.

Figure 6.4 shows a pie chart illustrating the test subjects' habits regarding mobile video watching. Although the widespread use of smartphones capable of playing video, the vast majority of respondents (41 of the 57) never watched a video via their mobile phone and only a minority of them (7 of the 57) watches mobile videos on a daily to weekly basis. Reasons for this limited usage of mobile video might be the high expenses of the cellular data transfer (in Belgium), and the battery consumption associated with the video playback.

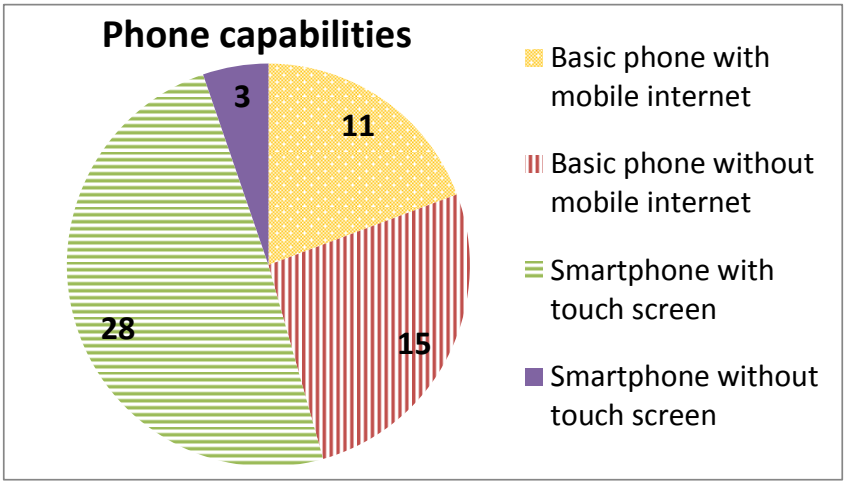


Figure 6.3: Pie chart showing the capabilities of the mobile phones that the test subjects own

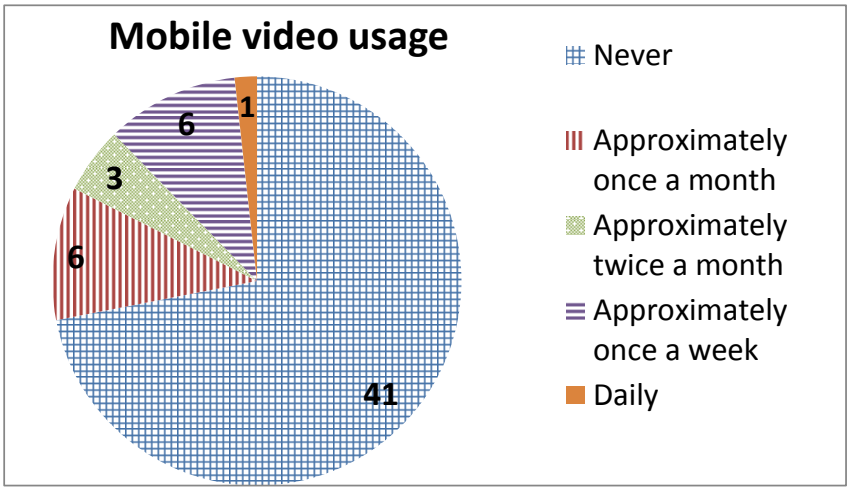


Figure 6.4: Pie chart showing the test subjects' habits regarding mobile video watching

The first open question was answered by almost one third of the participants (i.e., the test subjects who had prior experience with mobile video watching). These test subjects indicated which aspects and features that they think are important in view of having a good experience while watching a video on a mobile phone. The answers were coded in broader categories and counted. The figures mentioned here represent a percentage of the number of entries to this question. In general, the most important aspects that were mentioned are the general video

quality (22.7%), fluidity during the playback of the video (22.7%) and the audio quality (11.3%). Other aspects mentioned include the screen size and resolution, the absence of distortions, and the loading time of the video. Looking at the aspects that were mentioned first, the most important aspects are fluidity, video quality, and fast loading of the video. As these aspects were mentioned first, we can assume that they are so-called 'top of mind' for several of the test subjects and thus more important.

The answers to the second open question, which was inquiring after the expectations of the test subjects who had no prior experience with mobile video watching, were more diverse. More specifically, they were asked to indicate possible influencing factors, aspects of which they expected that they would influence their personal experience. Most mentioned in this respect are the loading speed (17.0%): fast loading of the video is expected to contribute positively to the experience, the screen properties (16.0%): the screen should be bright, big enough, and have the right resolution for the content, the quality of the audio and video (17.0%), and the fluidity of the video while playing (9.6%): there should be as little buffering interruptions as possible. Also mentioned several times are: synchronization of audio and video (3.2%), absence of distortions (4.3%), video player characteristics (4.3%), viewing context (7.4%), content (8.5%): the likeability of the content but also the availability of content, reliability of the internet connection (5.3%), and associated costs (6.4%). One participant also mentioned the battery of the device.

Finally, all the aspects of Table 6.5 are evaluated by the test subjects in terms of importance in order to have a good experience during mobile video watching. The second column of Table 6.5 lists the averages (arithmetic mean) of the test subjects' ratings obtained before the actual video experiment whereas the third column shows the averages of the ratings gathered after the experiment. All aspects are evaluated as important to have a good experience during video watching. "The fluidity of the image during video playback" and "the synchronization of image and sound during video playback" received the highest ratings.

The third column of Table 6.5 shows how test subjects evaluated the listed aspects immediately after the video experiment. Possibly influenced by the variable quality of the video sessions during the experiment, test subjects slightly changed their assessment compared to their initial ratings. The bold values in Table 6.5 indicate a significant difference between the ratings that test subjects gave before the experiment and the ratings specified by these test subjects after the experiment for the aspects mentioned in the first column. The significance of these differences is determined by the Wilcoxon signed rank test at a significance level of 0.05. The Wilcoxon signed rank test is a non-parametric statistical hypothesis test used to compare two related samples or repeated measurements on a single sample to assess whether their population means differ [9]. In this case, the two subjective ratings (before and after the experiment) originating from the same test subject are

compared. Given the discrete values of the test subjects' ratings, we opted for the non-parametric Wilcoxon signed rank test. However, the parametric counterpart of Wilcoxon's test, i.e., the statistical T-test, identifies the same statistical significant differences.

Aspect of the video	Average rating before the experiment	Average rating after the experiment
The content of the video	4.2	4.0
The technical quality of the video	**4.0**	**4.3**
The lack of distortions in the image during video playback	**4.3**	**4.5**
The fluidity of the image during video playback	**4.6**	**4.8**
The lack of distortions in the sound during video playback	4.3	4.5
The fluidity of the sound during video playback	4.4	4.6
Synchronization of image and sound during video playback	4.5	4.4
The loading speed of the video	**3.9**	**3.6**
The readability of text on the screen during video playback	3.8	3.6
The sharpness of the image during video playback	3.9	4.1
Other aspects: ...(to be completed by the test subject)	/	/

Table 6.5: Aspects of the video that test subjects had to evaluate in terms of importance in order to have a good experience during mobile video watching

Especially for those features that were impacted by the different connection types and video qualities, the differences could point to the adjustment of the test subjects' expectations and evaluation, based on their previous experience (i.e., during the test). For future research, it would be interesting to investigate how different levels of expectations (e.g., not met, met, exceeded) relate to specific quality levels and acceptability thresholds, how the current experience help to form or adjust those expectations, and how the expectations evolve over time due to subsequent experiences.

As mentioned in Section 6.2.2.3, after the experiment there was an additional open question asking the participants to prioritize the three aspects that according to them are most important in view of having a good QoE in the context of mobile video watching. When looking at all dimensions that were mentioned, most important are the fluidity of audio, the fluidity of audio and video in general, and the synchronization of audio and video. Additionally, the absence of distortions, the content, sharpness of the video image, and the loading time are considered to be of high importance. When we zoom in on the aspects that were mentioned first (highest priority), the most important aspects are the fluidity (both in general and of audio in particular), the absence of distortions, and the content.

To summarize, after the experiment test subjects attach significantly more importance to the technical quality of the video, the lack of distortions in the image, and the fluidity of the image during video playback. This increased importance may be due to the fact that some test subjects assess these technical aspects as unacceptable for some video sessions of the experiment. In contrast, the loading speed of the video is evaluated as less important after the experiment. This might be because test subjects assess the loading times during the experiment as acceptable and attach less importance to a short loading time than to a fluent playback of the video.

6.3.2 Objective measures

Table 6.6 shows the technical details regarding the video rebufferings and loading time, which are logged during each video playback. Although the loading time is limited to a few seconds for all connection types and quality versions of the video source, the median shows some characteristic differences for the six cases. Low-bandwidth connections induce longer loading times than medium- or high-bandwidth connections. As expected, the fastest loading times are measured for videos transmitted over a high-bandwidth connection and the high-quality videos require higher bit rates thereby causing longer loading times compared to low-quality video sources.

As mentioned earlier, bandwidth limitations can introduce interruptions during video playback due to rebufferings. However, the number of rebufferings and the point in time when such a rebuffering occurs, i.e., the rebuffering pattern, is non-trivial due to a number of interactions and correlations on several layers of the ISO/OSI stack [10]. The streaming server might implement flow control on the application layer; TCP implements flow control on the transport layer; the video player implementation (the built-in Android player in this experiment) tries to overcome interruptions by means of a video buffer; and the videos are encoded with variable bit rates. Still, differences in the rebuffer times are noticeable between low- and high-bandwidth connections as well as between low- and high-quality video sources. The median as well as the maximum of the measured rebuffer times are slightly higher for low-bandwidth connections and high-quality video sources compared to respectively high-bandwidth connections and low-quality video sources.

Table 6.6 illustrates that only a small number of rebufferings is required (median = 1) if a high-bandwidth connection is used or if a low-quality video source is transmitted over a medium-bandwidth connection. In these cases, the network connection provides sufficient throughput to transmit the video and prevent interruptions during video playback. For most video scenes, also a low-bandwidth connection provides sufficient throughput to transmit the low-quality video source.

	Low-quality source Low bandwidth	High-quality source Low bandwidth	Low-quality source Medium bandwidth	High-quality source Medium bandwidth	Low-quality source High bandwidth	High-quality source High bandwidth
Median loading time (seconds)	5.7	6.4	3.0	4.3	1.7	1.9
Median of a single rebuffer time (seconds)	0.9	1.0	0.7	1.0	0.7	0.7
Maximum single rebuffer time (seconds)	6.4	9.0	6.4	8.7	5.5	5.9
Median of the number of rebufferings	3	75	1	41	1	1
Standard deviation of the number of rebufferings	2.8	64.4	2.5	43.5	1.1	0.7
Median of the loading + total rebuffer time (seconds)	8.9	85.2	4.6	48.2	2.2	2.9
Standard deviation of the loading + total rebuffer time (seconds)	4.1	68.2	3.3	46.6	1.9	1.6

Table 6.6: Details about the measured rebuffering and loading times for the different connection types (low, medium, or high bandwidth) and quality versions of the video source (low or high quality)

However, peaks in the (variable) bit rate of the video may occasionally introduce rebufferings, which explains why the median of the number of rebufferings is 3 in this case.

On the other hand, the throughput obtained by using a low-bandwidth connection is insufficient for transmitting high-quality video sources fluently. This is confirmed by Table 6.6, which shows a large difference in the number of rebufferings for the high-quality video sources transmitted over a low-bandwidth connection compared to the other cases (e.g., the median of the number of rebufferings is 75 for high-quality video sources transmitted over a low-bandwidth connection). Also a medium-bandwidth connection provides insufficient throughput to transmit a high-quality video source without requiring rebuffering interruptions during playback (median of 41 rebufferings). Peaks in the video bit rate sometimes ex-

ceed the available network throughput. Still, the higher throughput of the medium-bandwidth connection compared to the throughput of the low-bandwidth connection reduces the (median of the) number of rebufferings by about half. Table 6.6 also shows that the standard deviation of the number of rebufferings during video playback is relatively high. Noise in the communication channel and the variable bit rate of the different videos result in a varying number of rebufferings for each combination of connection type and source quality. Therefore, the influence of these rebuffering interruptions on the different measures of QoE is investigated in Section 6.3.5.

In general, the period that video playback is interrupted by a rebuffering is quite short. Many interruptions last only a few hundred milliseconds and are hardly noticeable for the test subjects; (the median of this rebuffer time is 1 second or less, depending on the video quality and connection type). However, summing the (possible large amount of) rebufferings and the initial loading time of the video results in a substantial waiting time for the test subjects, ranging from 2.2 seconds for the most optimal solution to 85.2 seconds for the worst case. Therefore, we expect this waiting time together with the high frequency of rebufferings and the coupled video interruptions might deteriorate the quality of the user's experience significantly for some cases. The varying number of rebufferings for each combination of connection type and source quality results in a high standard deviation of the sum of the loading time and the total rebuffer time.

Given the high frequency of rebufferings and the short rebuffer times, the user's QoE might be improved by enlarging the buffer size thereby increasing the rebuffer times but reducing the frequency of rebufferings. However, since the built-in media player of the Android OS was used, changing the frequency of rebufferings or the buffer size was not possible in this experiment.

6.3.3 Subjective measures

We first take a closer look at the evaluation of the content and technical quality for the different technical scenarios. The histogram of Figure 6.5 visualizes the number of ratings gathered for each possible answer (going from 1 = Bad to 5 = Excellent) to the question in Table 6.4 regarding the content (question one). These subjective content evaluations provided by the participants of the experiment are partitioned according to the connection type and the quality of the video source. The large number of positive ratings (4 = Good or 5 = Excellent) indicates that most test subjects appreciate the content of the video experiment.

Moreover, the histogram illustrates that videos sent over a high-bandwidth connection received almost no negative evaluations regarding the content whereas video sessions using a medium- or low-bandwidth connection received a considerable number of negative assessments. Especially the content of video sessions in which a high-quality video is sent over a low-bandwidth connection (highQ

lowB) is poorly evaluated. So, the video sessions that suffered from the most rebufferings due to insufficient throughput of the network connection received the worst evaluation regarding the video content. In this scenario, more than 28% of the content ratings are negative, i.e., '1 = Bad' or '2 = Poor' (Figure 6.5). Also the high-quality video sources transmitted over a medium-bandwidth connection (highQ mediumB), which are also characterized by a lot of rebufferings, received a considerable number of negative evaluations regarding the content (12% of the content ratings are '1 = Bad' or '2 = Poor'). Finally, 14% of the low-quality video sources transmitted over a medium-bandwidth connection received a content rating of '1 = Bad' or '2 = Poor'. On the other hand, less than 4% of the video content that is transmitted over a high-bandwidth connection is negatively evaluated by the test subjects (i.e., received a rating of '1 = Bad' or '2 = Poor').

This difference in content appreciation, which is unlikely due to coincidence, indicates an effect of the technical quality of the video playback (and the coupled rebuffering interruptions) on the subjective evaluation of the content of the video. This finding is confirmed by the results of the experiment of Chapter 8 and our research regarding the influence of QoE on rating behaviour in recommender systems [11], which state that the user's subjective evaluation of the content is a combination of the user's preferences regarding the content and the subjective evaluation of the technical quality of the video.

Figure 6.6 shows the histogram of the ratings evaluating the technical quality partitioned according to the connection type and the quality of the video source. This histogram visualizes the test subjects' answers concerning question two of Table 6.4. High-quality video sent over a low-bandwidth connection (highQ lowB) received the worst evaluation from the test subjects as the majority of these sessions (71%) are evaluated as 'bad' or 'poor' on the technical quality. The reason for this poor evaluation is the high number of rebufferings and the coupled playback interruptions due to the low-bandwidth connection, as indicated in Table 6.6. Transmitting such a high-quality video over a medium-bandwidth connection (highQ mediumB), decreases the number of rebufferings by approximately 50% but still results in a suboptimal technical quality, as indicated by the considerable number of videos evaluated as 'bad' or 'poor' technical quality. However, the majority of the test subjects assesses the quality of these video sessions as 'fair' and the evaluations are roughly equally divided between positive and negative.

These two scenarios (high-quality video that is transmitted over a low- or medium-bandwidth connection) are the only scenarios which introduce a large number of rebufferings during video playback. Accordingly, only these scenarios received a considerable number of very negative evaluations (1 = Bad) regarding the technical quality from the test subjects. Other scenarios, in which video

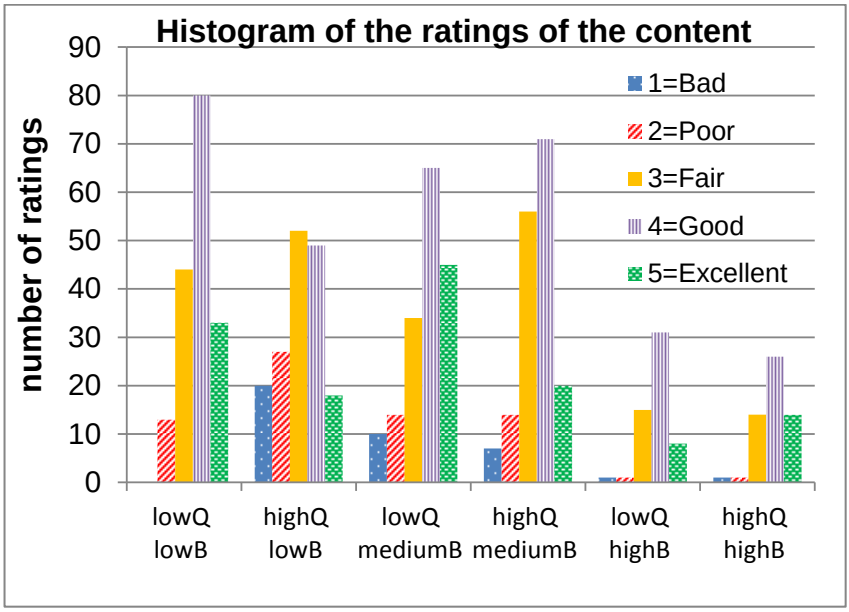


Figure 6.5: Histogram of the test subjects' ratings evaluating the content according to the connection type (low, medium, or high bandwidth (B)) and the quality (Q) of the video source (low or high). 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent

playback is not or only a few times interrupted, receive in general only neutral or positive evaluations on the technical quality.

Transmitting a low-quality video over a low-bandwidth connection (lowQ lowB) requires an acceptable number of rebufferings (the median of the number of rebufferings is 3 in Table 6.6). This almost fluent video-playback, together with the low bit rate and resolution induces a 'fair' evaluation of the technical quality (Figure 6.6), without many extreme positive or negative evaluations. Transmitting such a low-quality video over a medium-bandwidth connection (lowQ mediumB) reduces the number of rebufferings, which is reflected in more positive evaluations.

Video sessions using a high-bandwidth connection (lowQ highB and highQ highB) induce no or a very limited number of rebufferings thereby obtaining a very positive evaluation of the technical quality. E.g., 56% of the low-quality video sources and 97% of the high-quality video sources transmitted over a high-bandwidth connection received a rating of '4 = good' or '5 = excellent' on the technical quality. As expected, the best results are obtained by transmitting a high-quality video over a high-bandwidth connection (highQ highB). The high resolution and bit rate together with the fluent video playback convince test subjects to evaluate the technical quality of these sessions as 'good' or 'excellent'.

This histogram indicates that transmitting high-quality video sources is only useful if enough bandwidth is available. For low or medium-bandwidth connections, the best option is to reduce the bit rate and resolution of the video source and to transmit these low-quality video sources thereby preventing rebuffering interruptions as much as possible.

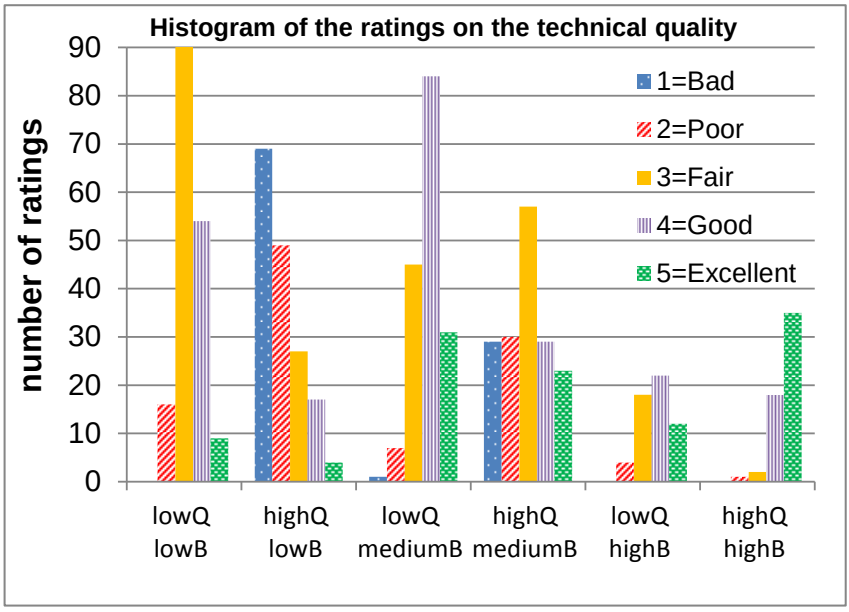


Figure 6.6: Histogram of the test subjects' ratings evaluating the technical quality of the video according to the connection type (low, medium, or high bandwidth (B)) and the quality (Q) of the video source (low or high). 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent

As indicated in Table 6.7, the subjective evaluations of the technical aspects of the video are positively correlated to each other. All these Pearson correlations are significant at the level of 0.99 ($p < 0.01$). The quality of the video source and the available bandwidth of the communication channel are the common factors that influence the subjective evaluations of all these technical aspects of the video. As a result, the histograms of all these subjective evaluations have similar distributions and do not reveal additional insight; so they are omitted in this dissertation.

Table 6.7 also shows that the subjective evaluations of the technical quality and the overall experience are positively correlated (based on Pearson's metric), proving the consistency of these general subjective parameters. Also the acceptability of the technical quality is in line with these subjective parameters: if the quality is evaluated as 'acceptable', the average (arithmetic mean) of the subjec-

tive evaluations of the technical quality and overall experience are respectively 3.8 and 3.5 whereas for video sessions assessed as 'unacceptable', these averages are respectively 2.1 and 2.0.

	Content	Technical quality	Interruptions	Fluidity	Loading	Experience
Content	1.000	0.300	0.245	0.272	0.298	0.593
Technical quality	0.300	1.000	0.744	0.737	0.507	0.595
Interruptions	0.245	0.744	1.000	0.826	0.563	0.598
Fluidity	0.272	0.737	0.826	1.000	0.562	0.625
Loading	0.298	0.507	0.563	0.562	1.000	0.495
Experience	0.593	0.595	0.598	0.625	0.495	1.000

Table 6.7: Correlations between the subjective evaluations

To quantify the influence of the objective parameters of Table 6.6 on the subjective evaluations of the video obtained via the questionnaire (Table 6.4), the Pearson correlation, ρ, is calculated to measure of how well the objective and subjective parameters are related. Table 6.8 lists the correlations between the parameters of the video session (i.e., the number of rebufferings, the waiting time, which is defined as the sum of the loading time and rebuffer times, and the mean RTT) and the subjective evaluations regarding the aspects of Table 6.4. All these Pearson correlations are significant at the level of 0.99 ($p < 0.01$). The results show a strong negative correlation between on the one hand the subjective evaluation of the user's experience and the ratings related to the quality of the video session (Technical quality, Interruptions, Fluidity, and Loading), and on the other hand the number of rebufferings and the time that these rebufferings require (together with the initial loading time of the video). This confirms that the subjective evaluation of the video quality and the coupled QoE are strongly influenced by the duration and amount of interruptions during video playback. A negative correlation is also observed between the mean RTT and the evaluation of the experience ($\rho = -0.345$) as well as the ratings with regard to the subjective quality of the video session: technical quality, interruptions, fluidity, and loading ($\rho \leq -0.393$). So, high round-trip delay times may have a negative influence on the users' QoE and the subjectively-observed video quality.

Noteworthy is the significant negative correlation between on the one hand the users' evaluation of the content and on the other hand the number of rebufferings ($\rho = -0.203$), the time that these rebufferings require ($\rho = -0.208$), and the mean RTT ($\rho = -0.191$). Although users were supposed to evaluate the content of the video regardless of the loading time, rebuffering interruptions, network character-

istics, and technical quality of the video, content ratings are clearly influenced by these technical parameters.

	Number of rebufferings	Waiting time = loading time + rebuffer times	Mean RTT
Content	-0.203	-0.208	-0.191
Technical quality	-0.552	-0.562	-0.421
Interruptions	-0.613	-0.622	-0.393
Fluidity	-0.705	-0.712	-0.397
Loading	-0.452	-0.478	-0.441
Experience	-0.510	-0.518	-0.345

Table 6.8: Correlations between the objective parameters of the video session and the subjective evaluations

6.3.4 Subjective technical quality and overall experience

The correlations of Table 6.8 confirm the influence of the objective parameters of the video sessions on the subjectively-observed video quality. Subsequently, to investigate which technical scenarios show a significant difference in subjective technical quality and overall experience, they are compared pairwise via a statistical test.

Since the test subjects' evaluations are discrete values, the six video scenarios were compared using the Wilcoxon rank-sum test, as also used in chapter 3. This way, the subjective ratings of the technical quality and of the overall experience (dependent variables) were compared using the different technical combinations (connection type and source quality) as the grouping variable (independent variable).

Table 6.9 shows the results of this Wilcoxon rank-sum test. The second column specifies which two scenarios (characterized by the connection type and the quality of the video source) are tested for a significant difference in the mean rating specified by the user. For reference purpose, each of these tests received a sequence number in the first column of the table. The third and fifth column show the point estimation of the difference between the mean values of the two scenarios (first scenario minus second scenario) for respectively the technical quality and the overall experience. The standard error on this point estimation is indicated between brackets. The p-value (of the fourth and sixth column) is an indication for the significance of the difference between the two scenarios. If the p-value is below 0.05, the evaluations of the scenarios are considered as significantly different.

	Difference of two scenarios	Difference in mean technical quality (Std)	P-value technical quality	Difference in mean QoE (Std)	P-value QoE
1	lowQ lowB - lowQ mediumB	-0.486 (0.105)	$< 10^{-4}$	-0.105 (0.106)	0.2191
2	lowQ lowB - lowQ highB	-0.421 (0.149)	0.0008	-0.189 (0.151)	0.1382
3	highQ lowB - highQ mediumB	-0.899 (0.106)	$< 10^{-4}$	-0.990 (0.107)	$< 10^{-4}$
4	highQ lowB - highQ highB	-2.530 (0.149)	$< 10^{-4}$	-2.190 (0.151)	$< 10^{-4}$
5	highQ mediumB - highQ highB	-1.630 (0.149)	$< 10^{-4}$	-1.200 (0.151)	$< 10^{-4}$
6	lowQ mediumB - lowQ highB	0.066 (0.149)	0.5793	-0.0833 (0.151)	0.6554
7	highQ lowB - lowQ lowB	-1.310 (0.105)	$< 10^{-4}$	-1.520 (0.107)	$< 10^{-4}$
8	highQ lowB - lowQ mediumB	-1.790 (0.106)	$< 10^{-4}$	-1.630 (0.107)	$< 10^{-4}$
9	highQ lowB - lowQ highB	-1.730 (0.149)	$< 10^{-4}$	-1.710 (0.151)	$< 10^{-4}$
10	highQ mediumB - lowQ mediumB	-0.893 (0.105)	$< 10^{-4}$	-0.637 (0.107)	$< 10^{-4}$
11	highQ mediumB - lowQ highB	-0.827 (0.149)	$< 10^{-4}$	-0.720 (0.151)	$< 10^{-4}$
12	highQ mediumB - lowQ lowB	-0.407 (0.105)	0.0008	-0.532 (0.106)	$< 10^{-4}$
13	lowQ lowB - highQ highB	-1.220 (0.149)	$< 10^{-4}$	-0.671 (0.151)	$< 10^{-4}$
14	lowQ mediumB - highQ highB	-0.738 (0.149)	$< 10^{-4}$	-0.565 (0.151)	0.0001
15	highQ highB - lowQ highB	0.804 (0.183)	$< 10^{-4}$	0.482 (0.185)	0.0009

Table 6.9: Results of the Wilcoxon rank-sum test performed on the subjective evaluations of the technical quality and the overall experience

Test 1, 2, 3, 4, and 5 compare the subjectively-observed quality and overall experience of video sessions using two network connections with a different bandwidth. For each of these tests, the quality of the video source is identical for the two scenarios (low quality for test 1 and 2; high quality for test 3, 4 and 5), whereas the bandwidth of the connection in the second scenario is higher than the bandwidth of the connection in the first scenario. The significant differences in subjective technical quality and overall experience as well as the negative values of the point estimations of these differences prove that users notice the more fluent video playback (i.e., less and shorter rebufferings as well as a shorter loading time) if a higher bandwidth is available for transmission. Only for test 1 and 2, the difference in overall experience was not found to be significant.

For test 4, the point estimation of the difference between the mean values of the observed technical quality and between the mean values of the overall experience is respectively -2.530 and -2.190. High-quality video sources that are sent over a

low-bandwidth connection are characterized by a large number of rebufferings and receive therefore a low evaluation. High-quality video sources transmitted over a high-bandwidth connection on the other hand, deliver a perfect image quality and require no or a very limited number of rebufferings during playback. Therefore, the biggest difference in subjective technical quality and overall experience is measured for these two extreme situations.

Also test 6 compares the observed technical quality and overall experience of video sessions using two network connections with a different bandwidth. However, this test shows no significant differences if a high-bandwidth connection is used instead of a medium-bandwidth connection for the transmission of a low-quality video. Since a medium-bandwidth connection provides already sufficient throughput for transmitting a low-quality video fluently, switching to a high bandwidth connection brings no further improvement in the observed technical quality or overall experience.

Test 7 shows a significant difference in observed technical quality and overall experience between high-quality and low-quality video sources that are transmitted over a low-bandwidth connection. The negative values of the point estimations of the differences between the mean values indicate that users provide a better evaluation for the low-quality video source. The reason for this is the high number of rebufferings that users experience if a high-quality video source is transmitted over a low-bandwidth connection. This indicates that in this case users prefer a more fluent playback of the video above a higher resolution, frame rate, and bit rate. So, content providers can optimize the subjectively-observed quality and overall experience of the video session by adapting the resolution, frame rate, and bite rate of the video depending on the available bandwidth. E.g., if the available bandwidth of the data connection is low, the best option is to transmit a low-quality video instead of a high-quality video to the end-user.

Test 8 and 9 further compare the playback of a high-quality video source using a low-bandwidth connection, which introduces a large number of rebufferings, with the (almost) fluent playback of low-quality video. Whereas test 7 uses a low-bandwidth connection for the transmission of the low-quality video, thereby causing an acceptable number of rebufferings (median=3), test 8 and 9 transmit the low-quality video using respectively a medium- and high-bandwidth connection, which require no or only a very limited number of rebuffering interruptions (median=1). This further decrease in the number of rebufferings leads to a better technical quality and overall experience. This is reflected in the higher absolute values of the point estimations of the differences between the mean values of the two scenarios.

Test 10, 11 and 12 compare the subjective quality and overall experience of a high-quality video source transmitted over a medium-bandwidth connection with low-quality video transmitted over the three connection types. According to the

results of test 10, the playback of a low-quality video source results in a better subjective quality and overall experience than the playback of a high-quality video source if the transmission channel has a medium-bandwidth connection. Again, the number of rebufferings and the coupled playback interruptions are the reasons why users prefer a low-quality video above a high-quality video if a medium-bandwidth connection is available. So also in this scenario, users prefer a fluent playback of their video, even if this means that they have to sacrifice resolution and frame rate.

In test 11, a high-bandwidth connection is used as communication channel for the low-quality video in contrast to the medium-bandwidth connection of test 10. This high-bandwidth connection causes no further improvement since a medium-bandwidth connection offers already sufficient throughput for transmitting the low-quality video without introducing too many rebufferings. Test 12 shows another interesting result. Video sessions using a low-quality video source and a low-bandwidth connection are significantly better assessed than video sessions based on a high-quality video source and a medium-bandwidth connection. Since the throughput of the medium-bandwidth connection is still insufficient for transmitting high-quality videos and thereby requires too much rebufferings, this test confirms the users' preference for fluent video playback above high-quality video sources.

Test 13 compares two opposite cases: low-quality video over a low-bandwidth connection against high-quality video over a high-bandwidth connection. As expected, the high-quality video using a high-bandwidth connection receives an assessment that is much better than the low-quality video sent over a low-bandwidth connection. Although the estimated differences between the mean ratings for these two scenarios are very significant (-1.220 for technical quality and -0.671 for overall experience), these are not the biggest differences that were encountered in the experiment. (Test 4 showed the biggest differences between the mean ratings for the two scenarios.)

Lastly, test 14 and 15 represent cases in which sufficient bandwidth is available for video transmission and the number of video rebufferings remains limited. In test 14, sending a low-quality video over a medium-bandwidth connection is compared with the transmission of a high-quality video over a high-bandwidth connection. Since video playback is fluent for both cases, the only discriminating factor is the quality of the video source. Therefore, the high-quality video (which is sent over a high-bandwidth connection) is evaluated better than the low-quality video (which uses the medium-bandwidth connection). Finally, test 15 compares low-quality and high-quality video sources which are both transmitted over a high-bandwidth connection. Since this connection provides enough throughput for both quality versions, the difference in observed technical quality and overall experience is merely based on the difference in the quality of the video sources. As

expected, the high-quality video source is assessed significantly higher than the low-quality video source.

The other subjective evaluations regarding the technical properties of the video (Interruptions, Fluidity, and Loading) show similar results, also pointing to the consistency of test subjects in their ratings. Almost every test reveals significant differences between the video scenarios. Even the evaluation of the content of the video shows to be significantly different for various couples of video scenarios. For example, the content of the high-quality video transmitted over a high-bandwidth connection is significantly better assessed than the content of high-quality video transmitted over a low-bandwidth connection according to a Wilcoxon rank-sum test. Again, this supports the assumption that the subjectively-observed technical quality and the overall experience are aspects that influence the subjective evaluation of the content.

6.3.5 Acceptability of the technical quality

Besides knowing which video scenarios receive a different evaluation regarding the technical quality and overall experience, it is essential to identify the video scenarios with a technical quality that is 'acceptable' according to the users. This means, video sessions have to be classified as 'acceptable quality' or 'unacceptable quality'. Therefore, test subjects, not informed about the source quality or connection type, were asked to evaluate the acceptability of the technical quality of each video during the experiment via the last question of Table 6.4. Table 6.10 summarizes this acceptability of the technical quality for the different video scenarios.

The high-quality video sources that are sent over a low-bandwidth connection and thereby require numerous rebufferings during video playback are in general evaluated as 'unacceptable'. The frequent rebufferings (median number = 75 (Table 6.6)) are experienced as annoying and often even intolerable, since only 7.8% of these video sessions are evaluated as 'acceptable'. If a medium-bandwidth connection is used to transmit such a high-quality video, still a considerable number of rebufferings is necessary during the video playback (median number = 41 (Table 6.6)). Despite this high number of rebuffering interruptions, the technical quality of 39.3% of these sessions is evaluated as 'acceptable'. The reason for this might be that users' expectations regarding the technical quality of mobile video services can be quite low, hereby expecting and accepting interruptions during video playback.

The sessions in which low-quality video is transmitted over a low-bandwidth connection undergo a limited number of rebufferings (median number = 3 (Table 6.6)). The combination of this low-quality video source and the small number of playback interruptions is accepted in 85.3% of the cases. Since low-quality video can fluently be transmitted over a high- or medium-bandwidth connection

without requiring rebufferings, the technical quality of these video sessions is almost always acceptable (in respectively 91.1% and 94.6% of the cases).

The highest acceptance rate (96.4%) of this experiment is measured for high-quality video sources, transmitted over a high-bandwidth connection. The fluent playback of this high-quality video provides the most optimal video rendering on the mobile device but requires a high throughput to prevent rebuffering interruptions.

	Total number of ratings	Number of acceptable sessions	Number of unacceptable sessions	Rate of acceptance (%)
High-quality source Low bandwidth	166	13	153	7.8
High-quality source Medium bandwidth	168	66	102	39.3
Low-quality source Low bandwidth	170	145	25	85.3
Low-quality source High bandwidth	56	51	5	91.1
Low-quality source Medium bandwidth	168	159	9	94.6
High-quality source High bandwidth	56	54	2	96.4

Table 6.10: Evaluation of the acceptability of the observed video quality for the different combinations of connection type and quality of the video source

To obtain a model to predict the acceptability of the technical quality of the video session, a (binary) logistic regression analysis was performed. Logistic regression is used to predict the probability of an event (in this case, the rejection of the video quality) by fitting data to a logistic curve [8]. In contrast to the analysis of the subjectively-observed technical quality and overall experience in Section 6.3.4, the acceptability of the technical quality is modelled via a logistic regression, because of the binary nature of this evaluation.

Because of the significant correlations (Table 6.8) between the subjective evaluations and the number of rebufferings during video playback, we opted for the number of rebufferings as a predictor variable (independent variable) and the acceptability of the technical quality is chosen as the dependent variable. The result of this logistic regression analysis is a model for the probability, p, that the user will not accept the quality of the video. The resulting equation 6.1 illustrates that the probability of an unacceptable quality increases as the number of rebuffer-

ings increases. In this equation, $NoRebuf$ stands for the number of rebufferings during the playback of a single video on the mobile device. A critical point is reached when the number of rebufferings is greater than 32, since the probability of an unacceptable quality is then higher than 50%. This number of rebufferings might seem quite high, but can be explained by the settings of the experiment. For video watching on a mobile phone, users might lower their expectations regarding the technical quality compared to video watching on a desktop computer or a television set, and thereby accept some playback failures such as rebuffering interruptions. Moreover, these rebuffering interruptions of the Android player are very short (i.e., a median duration between 0.7 and 1.0 seconds, depending on the connection and source quality) and sometimes even hardly noticeable. As a result, the technical quality of video sessions with ten or less rebufferings are in most cases (more than 80%) accepted by the users.

$$p = \frac{e^{-2.120+0.065\ NoRebuf}}{1 + e^{-2.120+0.065\ NoRebuf}} \tag{6.1}$$

The model of equation 6.1 is based on the subjective evaluations of the acceptability of the technical quality of 782 video sessions. The null deviance of this model is 1039 whereas the residual deviance is 564, which is smaller than the 95% quantile of the χ^2 distribution with 782 degrees of freedom i.e., $\chi^2(0.95, 782) = 848$. This statistical test confirms that the data (i.e., the subjective acceptability and the number of rebufferings) is distributed according to the proposed logistic regression model [8].

Figure 6.7 visualizes the result of this logistic regression analysis by plotting the probability of an unacceptable technical quality as a function of the number of rebufferings, which is varying from 0 to 100 (line diagram). Figure 6.7 also compares this logistic curve with the subjective evaluations of the acceptability, obtained through the questionnaire of the experiment (bar diagram). Therefore, the video sessions of the experiment are classified according to the measured number of rebufferings. Each of the video classes has a range of 10 units regarding the number of rebufferings. Next, the fraction of video sessions that are evaluated as 'unacceptable' during the experiment is calculated for each of these classes and visualized in Figure 6.7 as 'measured probability'. So the line diagram estimates the probability of an unacceptable quality based on the logistic regression analysis, whereas the bar diagram shows the fraction of videos that were evaluated as 'unacceptable' in the video experiment. The graph shows that the estimated probability is a good fit of the measured fraction of unacceptable videos, which is calculated based on the subjective evaluations. This is confirmed by the RMSE (Root Mean Square Error) of 0.18, which is calculated based on the difference between the predicted probability and the measured probability.

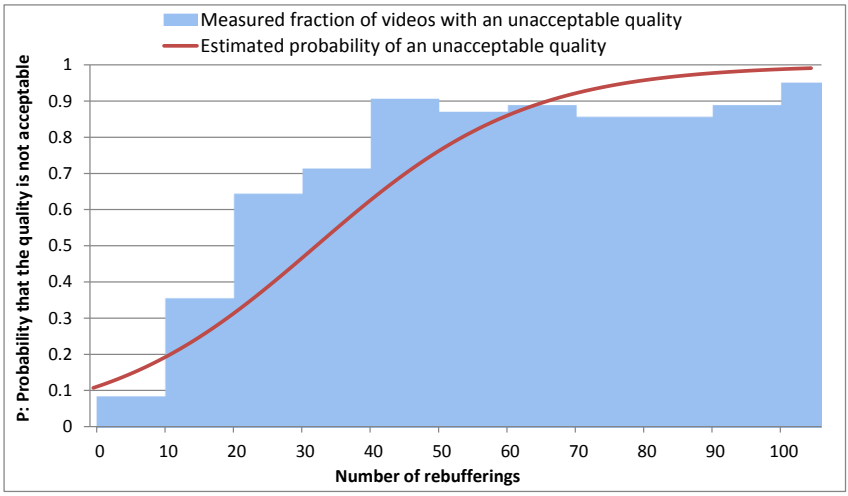


Figure 6.7: The probability that the technical quality of the video is not accepted by the user, as a function of the number of rebufferings during the video session

In order to take into account also the loading time of the video, this logistic regression analysis is repeated with the waiting time (i.e., the sum of the loading time and rebuffer times during video playback) as a predictor variable. Hence in this analysis, the waiting time is the independent variable and the acceptability of the technical quality is chosen as the dependent variable. The result is again a model for the probability, p, that the user will not accept the quality of the video. As expected, equation 6.2 shows that the probability of an unacceptable quality increases as the waiting time, denoted by $WaitTime$, increases. When the waiting time becomes more than 39 seconds, a critical point is reached and the probability of an unacceptable quality is then higher than 50%.

$$p = \frac{e^{-2.444+0.063\,WaitTime}}{1+e^{-2.444+0.063\,WaitTime}} \tag{6.2}$$

Just as in the analysis with the number of rebufferings as predictor variable, this model is based on 782 subjective evaluations of the acceptability of the technical quality of a video, and the null deviance is 1039. For this model the residual deviance is 547, which is smaller than the 95% quantile of the χ^2 distribution with 782 degrees of freedom i.e., $\chi^2(0.95, 782) = 848$. So, this statistical test [8] confirms that the subjective acceptability evolves as a function of the waiting time, according to the proposed logistic regression model of equation 6.2. The lower residual deviance of this model (i.e., 547) compared to the residual deviance of the model based on the number of rebufferings (i.e., 564) indicates that this model is a slightly better fit for the acceptability of the technical quality.

Figure 6.8 visualizes the result of this logistic regression analysis by plotting the probability of an unacceptable technical quality as a function of the waiting time, which is varying from 0 to 130 seconds (line diagram). Just as in Figure 6.7, the logistic curve is compared with the subjective evaluations of the acceptability, obtained through the questionnaire of the experiment (bar diagram in Figure 6.8). Therefore, the video sessions are classified according to the objective waiting time. Each of the video classes has a range of 10 seconds in waiting time. Subsequently, the fraction of video sessions that are evaluated as 'unacceptable' by the test subjects is calculated for each of these classes and visualized in Figure 6.8 as 'measured probability'. The graph shows that the estimated probability (based on the logistic regression) is a good fit of the measured fraction of unacceptable videos, which is based on the subjective evaluations of the questionnaire. Moreover, the RMSE of 0.15 confirms the suitability of the model based on the waiting time and indicates that this model is even a slightly better fit for the acceptability than the model based on the number of rebufferings, which has an RMSE of 0.18.

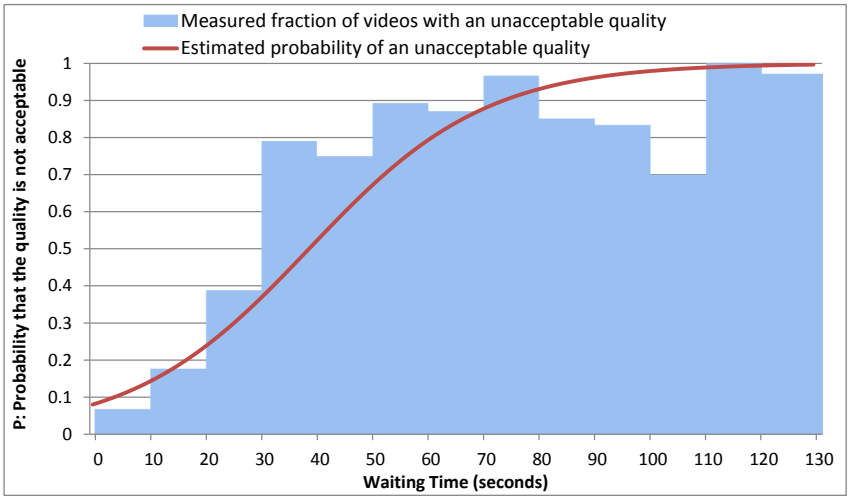


Figure 6.8: The probability that the technical quality of the video is not accepted by the user, as a function of the waiting time during the video session

6.4 Conclusions

This chapter presented the results from an exploratory study on the (acceptability of the) technical quality of mobile video and the overall experience of users during mobile video watching. The controlled environment of this experiment allowed us to manipulate the bandwidth of the data connection used to transfer the videos

to the mobile device. Three different connection types (low, medium, and high bandwidth) were combined with two levels of visual quality of the video source.

Subjective quality assessments, obtained via a questionnaire, showed to be highly correlated with the measured objective parameters of the video session, such as the number of rebufferings, the rebuffer time, and the loading time of the video. Although video interruptions due to rebufferings are experienced as disturbing, users accept a (limited) number of these rebufferings in a mobile context. Furthermore, the subjective evaluations of the video quality indicated that the test subjects of our experiment preferred a fluent playback of the video above a higher resolution, frame rate, and bit rate. In comparison with the fluidity of the playback, the test subjects considered the loading time of the video as less critical for having a good experience.

Based on the subjective evaluations of the users and the objective parameters of the video sessions, this study models the acceptability of the quality of a mobile video session. Due to the significant influence of a few objective parameters on the subjective experience of the user, an accurate prediction of the user's acceptance can be made based on the number of rebufferings or the waiting time during video playback (i.e., the sum of the loading time and the rebuffer times). Although only a limited number of objective parameters are used in the resulting models, the evaluation showed that the models are a good fit of the data obtained through the questionnaire. These models estimate the probability that users will accept the quality of a mobile video session as a function of the number of rebufferings or the waiting time during video playback.

The more rebufferings the lower the probability that users consider the technical quality as 'acceptable'. Still, mobile video sessions with less than 10 short rebufferings (with a median duration of 1 second) are in most cases (more than 80%) evaluated as 'acceptable'. In contrast, if more than 50 of these short rebufferings occur during playback, the technical quality of the video session is typically (more than 75%) not acceptable. The probability of acceptance can also be expressed in terms of the waiting time during playback. Video sessions with a waiting time below 20 seconds have a high probability (more than 75%) to be accepted by the user, whereas sessions with more than 60 seconds of waiting time are in general (more than 75%) evaluated as 'not acceptable'. This proposed QoE model enables operators to fix performance targets in terms of human perception.

Future research should however seek to validate these findings, not only in controlled research settings but also in more ecologically valid[1] usage contexts. The setup of a complementary living lab or field study, in which the influence of physical as well as social contextual factors can be more closely investigated would be a first step towards a more natural usage environment. Secondly, to take into account

[1]The ecological validity of a study refers to the methods, materials, and settings of the study that must approximate the real-world that is being examined.

the influence of temporal dimensions and effects, a study with a longer time frame (e.g., one to several weeks) could be set up. Finally, it would be very relevant to also look at other types of mobile devices enabling mobile video watching (for instance smartphones vs. tablets) to see if test subjects adjust their expectations and acceptability thresholds depending on the technical context (e.g., screen size). As such, it could be further investigated which additional factors might affect users' overall experience and their acceptance or refusal of the produced quality as well as how these factors can be taken into account in order to optimize the experience.

References

[1] S. Jumisko-Pyykkö, V. K. Malamal Vadakital, and M. M. Hannuksela. *Acceptance threshold: A bidimensional research method for user-oriented quality evaluation studies*. International Journal of Digital Multimedia Broadcasting, 2008:1–20, 2008.

[2] Speedtest.net. *The global broadband speed test*, 2013. Online available at `http://www.speedtest.net/`.

[3] *UMTS network capacity planning*. Technical report, UMTS World, 2003. Online available at `http://www.umtsworld.com/technology/capacity.htm`.

[4] *H.264 : Advanced video coding for generic audiovisual services*. Technical report, ITU-T, International Telecommunication Union, January 2012. Online available at `http://www.itu.int/rec/T-REC-H.264-201201-I/en`.

[5] G. Combs. *The world's foremost network protocol analyzer*, 2012. Online available at `http://www.wireshark.org/`.

[6] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake. *Study of Rating Scales for Subjective Quality Assessment of High-Definition Video*. IEEE Transactions on Broadcasting, 57(1):1 –14, March 2011.

[7] *P.911: Subjective audiovisual quality assessment methods for multimedia applications*. Technical report, ITU-T, International Telecommunication Union, 1998. Online available at `http://www.itu.int/rec/T-REC-P.911-199812-I/en`.

[8] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. McGraw-Hill, fifth edition, 2005.

[9] J. J. Higgins. *Introduction to Modern Nonparametric Statistics.* Thomson Brooks/Cole, first edition, 2004.

[10] T. Hoßfeld, R. Schatz, T. Zinner, M. Seufert, and P. Tran-Gia. *Transport Protocol Influences on YouTube Videostreaming QoE*. Science research report series, University of Würzburg Institute of Computer, 2011. Online available at `http://www3.informatik.uni-wuerzburg.de/TR/tr482.pdf`.

[11] T. De Pessemier, K. De Moor, I. Ketykó, W. Joseph, L. De Marez, and L. Martens. *Investigating the influence of QoS on personal evaluation behaviour in a mobile context*. Multimedia Tools and Applications, 57:335–358, 2012.

7

QoE research in a living lab environment

7.1 Introduction

Notwithstanding the importance of Quality of Experience (QoE) during the design and development of mobile applications and services, QoE assessment is still challenging, especially in real-life (so called 'living lab') contexts. Despite the lack of absolute control over the operating conditions, it can be argued that research in more realistic settings yields a much higher ecological validity. Especially when focusing on mobile applications and services, research 'in context' may complement controlled laboratory testing. This chapter presents results from an exploratory study on QoE and more specifically, on the subjective evaluations of various quality aspects during video watching on a mobile device, in a realistic setting. The results of this study may contribute to the estimation of users' subjective evaluation of the quality during mobile video watching and to QoE optimization by dynamically altering the parameters that have the largest influence on this subjective evaluation.

7.2 Test setup

7.2.1 Goals of the study

The main aim of this study is to investigate QoE aspects related to mobile video watching in a living lab setting. More specifically, this chapter zooms in on usage patterns in a natural research context and on the subjective evaluation of high- and low-quality movie trailers that are transferred to a mobile device using two transmission protocols for video (i.e., RTP (real-time transport protocol) over UDP (User Datagram Protocol) and progressive download using HTTP(HyperText Transfer Protocol)). Via an exploratory experiment taking place in a realistic environment, objective and subjective QoE aspects are gathered, having in mind the end goal of obtaining a methodology for the multi-dimensional quantification of QoE during mobile video watching. Similar to the experiment in the controlled environment of Chapter 6, user feedback was collected by means of short questionnaires on the mobile device, combined with traditional pen and paper diaries. Although the gathering of immediate subjective user feedback may to some degree interfere with the experience of the user, it has the important advantage that the experience can take place in the natural user context while at the same time feedback can be collected on different levels without delay between the actual experience and the feedback on the experiences. The subjective evaluations regarding the general technical quality, perceived distortion, fluidity of the video, and loading speed are studied and the influence of the transmission protocol and video quality on these evaluations is analysed.

7.2.2 Procedure

For this experiment, the test subjects were asked to watch 28 pre-defined movie trailers (covering different genres) in their everyday life context (when and where they wanted), but within a time-span of one week (weekend included).

7.2.2.1 Phase 1: Instruction meetings

Before the actual test started for every test subject, instruction meetings were organized in groups of five test subjects. Because of the living lab environment, providing assistance during the experiment was more difficult than in the case of the controlled laboratory environment and therefore special attention was paid to the briefing of the test subjects. After some general information on how to switch on/off, use, charge the device etc., it was explained how to access the test application and how to select and watch the videos. Because test subjects had to evaluate each video via a small questionnaire on the device, it was also shown how to navigate from one question to the next and fill in the questionnaire using the touch

screen. At the end of the briefing session, every test subject was given a device, a diary, and an instruction leaflet with practical information, screenshots, and relevant instructions related to the grading scales and univocal interpretation of the questions.

7.2.2.2 Phase 2: mobile video watching in a living lab environment

Figure 7.1 shows the architecture of the video delivery system used in the living lab experiment, consisting of the client device (i.e., a smartphone running the video player), the video server offering the content, and the technical database storing the objective parameters and subjective evaluations. In contrast to the architecture of the controlled lab experiment (Figure 6.1), videos can be transmitted using RTP streaming or progressive download; and for this living lab experiment, WiFi networks as well as the commercial cellular data network of Proximus, a Belgian network operator, were used. Through the Proximus network, the client device is connected to a GPRS (General Packet Radio Service), EDGE (Enhanced Data rates for GSM Evolution), UMTS (Universal Mobile Telecommunications System), or HSPA (High Speed Packet Access) network depending on the location of the test subject. Besides, test subjects can opt to connect their device to a WiFi network that is available for them, e.g., at their work or at their home.

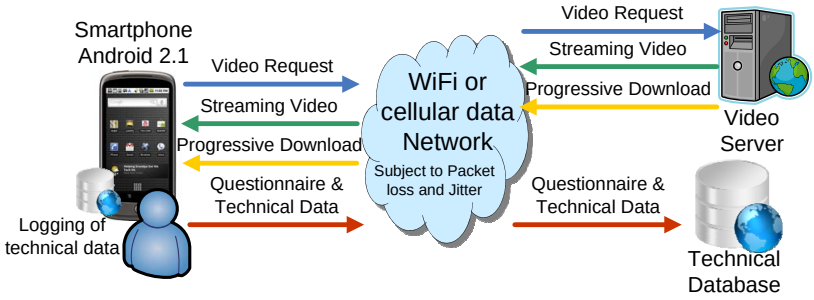


Figure 7.1: The architecture of the video delivery system used in the living lab experiment

All movie trailers used in the experiment were relatively short and had a duration between 2 and 3 minutes. The selected trailers covered different movie genres (such as comedy, drama, animation, action, and romance) and consisted of recent titles. To avoid boredom, the test subjects had to watch all 28 trailers only once during the experiment. The viewers were able to decide themselves in which order they watched the clips. The list consisted of 7 low-quality videos using RTP, 7 high-quality videos using RTP, 7 low-quality videos using progressive download, and 7 high-quality videos using progressive download. As explained in Chapter 5, both RTP and progressive download are often used for the transmission of

video content but have different characteristics in terms of possible influence on the user's experience.

These transmission protocols were combined with two video qualities in order to investigate their impact upon the user's quality evaluation. Table 7.1 summarizes the technical parameters of the two quality version of the mobile videos. All videos were coded with an average bit rate and resolution as specified in the table. The video list in the user interface was randomly mixed and the users were not informed about the different qualities and transmission protocols.

Because of the living lab environment and the possibly limited data throughput of cellular networks, some technical parameters of the video are chosen differently from the parameters of Table 6.2. In the living lab experiment, the average bit rate of the high-quality video is reduced to limit the packet loss (in case of RTP streaming) and rebuffering interruptions (in case of progressive download). Given the importance of the fluidity of the video (as expressed by the test subjects in the experiment of Chapter 6), the frame rate of the low-quality video was increased to obtain a more fluent playback of the video. Increasing the frame rate without increasing the bit rate too much was only possible by lowering the resolution of the low-quality video. In contrast to Table 6.2, in this experiment audio is coded with a different bit rate for the low- and high-quality video, resulting in a different audio quality for the two video types. This variation in the bit rate for the audio track strengthens the distinction between the low- and high-quality video.

Low Quality Video Source			
Audio		Video	
Codec	AAC LC 3	Codec	H.264/AVC
Average bit rate	32 kbit/s	Average bit rate	128 kbit/s
Maximum bit rate	44 kbit/s	Maximum bit rate	576 kbit/s
Channels	2	Resolution	142 x 80
Sampling frequency	44100 Hz	Frame rate	24 fps
High Quality Video Source			
Audio		Video	
Codec	AAC LC 3	Codec	H.264/AVC
Average bit rate	128 kbit/s	Average bit rate	384 kbit/s
Maximum bit rate	153 kbit/s	Maximum bit rate	922 kbit/s
Channels	2	Resolution	512 x 288
Sampling frequency	44100 Hz	Frame rate	24 fps

Table 7.1: Technical parameters of the mobile video used in the living lab environment

For this experiment, test subjects used Google Nexus One mobile phones, the same devices as used for the experiment in the controlled environment of Chap-

ter 6. The application that was used to watch the video has also the same interface as the application used for the experiment in Chapter 6. So the video selection mechanism and the questionnaire remains the same, as illustrated in Figure 6.2(a) and 6.2(b); only the video content and the questions were changed for the living lab experiment.

During the video watching, relevant objective video and network parameters were logged: video quality (resolution, frame rate, and bit rate), transmission protocol (RTP or progressive download), packet-loss rate for the audio and video track, the mean and maximum jitter for audio and video, network type(s) used for (a part of) the video transmission (e.g., UMTS, HSPA, GPRS), RSSI per network type (Received Signal Strength Indication), number of handovers (i.e., all kinds of radio cell reselections), and inter-system handovers (i.e., different data connection-type cell reselections, e.g., between UMTS and HSPA). In addition, a number of objective parameters concerning the video session and watching behaviour were registered: movement of the device (i.e., the GPS signal to track the mobility during the video watching), early interruption of the video (e.g., due to network disconnection), metadata about the video (ID, the coupled title, and duration) and the start and end of the session (timestamp). Table 7.2 provides an overview of these measured objective parameters with their unit, value, and sampling rate. Because of the living lab environment, additional parameters, such as the network type, are interesting to monitor compared to the experiment in the controlled environment (Table 6.3).

Figure 7.2 visualizes the locations of the observed data samples (based on GPS coordinates) grouped in clusters, which shows undoubtedly the true living lab environment in Flanders (Belgium) where most tests were conducted. The numbers in circles correspond to the number of video-watching sessions at that location.

In order to gather immediate and explicit user feedback after each watched video, six short questions concerning the content, general technical quality, fluidity of the video, loading speed, eventual distortions, and the user's physical context had to be answered on the device. Like in the experiment described in Chapter 6, these questions pop up on the screen after the video playback and users have to answer them before the next video can be played. In the briefing preceding the start of the experiment, the technical parameters that test subjects had to evaluate were explained to ensure an unambiguous interpretation. The questions with respect to the content, general technical quality, fluidity of the video, and loading speed, are the same as in Chapter 6 and the interpretation of these technical parameters was stated in Section 6.2.2.2.

Given the use of RTP as transmission protocol for several videos in this experiment, an additional question regarding the perceived distortion was added. Distortion was explained more broadly and different examples of possible distortions

Parameter	Unit	Value	Sampling Rate
Video ID	Integer	[0, ∞[	Each video
Video quality	State	{low, high}	Each video
Transport protocol	State	{RTP, prog. download}	Each video
Video packet loss rate	%	[0, 100]	Video RTP packets
Audio packet loss rate	%	[0, 100]	Audio RTP packets
Video jitter	Seconds	[0, ∞[	Video RTP packets
Audio jitter	Seconds	[0, ∞[	Audio RTP packets
GPRS percentage	%	[0, 100]	In each second
GPRS mean RSSI	dBm	[-113, -51]	In each second
EDGE percentage	%	[0, 100]	In each second
EDGE mean RSSI	dBm	[-113, -51]	In each second
UMTS percentage	%	[0, 100]	In each second
UMTS mean RSSI	dBm	[-113, -51]	In each second
HSPA percentage	%	[0, 100]	In each second
HSPA mean RSSI	dBm	[-113, -51]	In each second
WiFi percentage	%	[0, 100]	In each second
WiFi mean RSSI	dBm	[-113, -51]	In each second
Num. of handovers	Integer	[0, ∞[	In each second
Num. of inter-system handovers	Integer	[0, ∞[	In each second
Mobility	state	{indoor, no, slow, fast}	In each second
Percentage watched	%	[0, 100]	Each video
Start time	Epoch	Timestamp	Each video
Stop time	Epoch	Timestamp	Each video
Loading time	Seconds	[0, ∞[	Each video
Rebuffering times	Seconds	[0, ∞[	Each rebuffering
Mean RTT	Seconds	[0, ∞[	Each video

Table 7.2: The measured objective parameters of the video sessions in the living lab experiment

were given (e.g., blurriness, blockiness, ...). The perceived distortion was evaluated on a 5-level subjective quality evaluation scale, similar to the impairment rating scale of ITU [1], ranging from 5 (not perceptible) to 1 (perceptible and very annoying). Because of the living lab environment of this experiment, the last question on the device queried the test subjects about their physical context. Four options were selectable for the physical context of the user: 'on the move', 'at home', 'at work', or 'somewhere else'. In the case of selecting 'somewhere else', the user could specify his or her location. Table 7.3 lists the questions of this digital questionnaire and the used measurement scales.

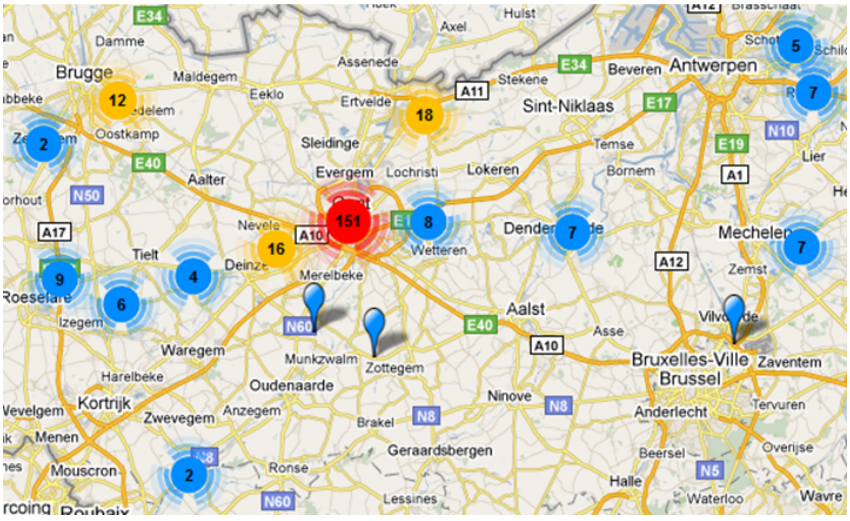


Figure 7.2: Location clusters of the user tests based on the GPS coordinates

Additionally, a traditional paper diary was completed by the test subjects immediately after playback: for every watched video, a diary sheet containing additional (open and closed) questions was filled in. The goal of this paper diary was to give users the opportunity to provide more detailed and qualitative feedback regarding the video session and their experience through some open questions. Since inputting text on mobile phones is difficult and tedious, mobile phones are not the optimal tool to gather detailed feedback. Therefore, we opted for an alternative feedback tool: a small paper diary that can also be used in case of technical problems with the device such as an application crash or a dead battery. Concerning the appreciation of the content, test subjects were firstly asked to indicate whether or not they would want to watch the entire movie and whether they had already seen it before. Secondly, they were asked to rate their general experience and to mention aspects that on the one hand influenced their experience (in a positive way as well as in a negative way) and on the other hand, that might help to improve the experience. The third question of the diary asked the test subjects whether other people were around the test subject during watching (in a radius of approximately 5 meter) and whether or not the presence of others was perceived as disturbing. Finally through the fourth question, test subjects had to indicate whether the overall technical quality of the video during the watching experience was a) acceptable in any context, b) acceptable but only in the context in which the test subject watched it, or c) not acceptable. Although each test subject watched each movie trailer in only one context, this question provides insights into the users' experiences and

Reference	Digital questions on the device	Possible answers
1. Content	How would you evaluate the content of the video?	5-point rating scale: 1 = Bad; 5 = Excellent
2. Technical Quality	How would you evaluate the technical quality of the video in general?	5-point rating scale: 1 = Bad; 5 = Excellent
3. Distortion	Did you perceive visual distortions in the video during playback?	5-point rating scale: 1 = Noticeable, very annoying; 5 = Not noticeable
4. Fluidity	How would you evaluate the fluidity of the video playback?	5-point rating scale: 1 = Bad; 5 = Excellent
5. Loading	How would you evaluate the loading speed of the video?	5-point rating scale: 1 = Bad; 5 = Excellent
6A. Location	Select your current location. I am ...	4 options: a) On the move, b) At home, c) At work, or d) Somewhere else.
6B. Location	(if (d) somewhere else) Where exactly are you?	Open Question

Table 7.3: The digital questions that were used to evaluate the video immediately after the playback in the living laboratory experiment, together with a reference to these questions and the possible answers

behaviour regarding video watching in different contexts. Table 7.4 lists the questions of this paper diary and the possible answers.

As already briefly mentioned above, the research design draws on two complementary voting interfaces because of the specific nature of the data that we wanted to collect. The 'on the device' voting interface is very suitable for collecting an immediate, in situ evaluation, as close to the experience as possible. As the short questionnaire on the device was part of the viewing protocol, we are sure that the test subjects rated the videos immediately after viewing. As a result, we were able to limit possible biases on the rating procedure due to memory errors or due to the time elapsed between the watching and the evaluation. At the same time, we deliberately aimed to limit the number of questions on the device as much as possible in order not to disrupt the user's natural flow when using the smartphone. However,

Reference	Paper diary questions	Possible answers
1A. Content Desirability	Please indicate whether or not you agree with the statement: "I would like to view the entire movie"?	5-point rating scale: 1 = Completely disagree; 5 = Completely agree
1B. Content Desirability	If you have already seen this movie before, please indicate by colouring the button.	Binary answer: Yes or No
2A. Experience	Please indicate on the scale below how you have experienced this viewing session. My overall experience was ...	5-point rating scale: 1 = Bad; 5 = Excellent
2B. Positive Aspects	Which aspects did you experience as positive during the viewing session?	Open question
2C. Negative Aspects	Which aspects did you experience as negative during the viewing session?	Open question
2D. Enhancing Aspects	Which aspects could enhance or improve your experience?	Open question
3A. Other People	Were other people in a radius of approximately 5 meter around during the viewing session? If so, how many?	No or Yes + number
3B. Other People	(If other people were in the immediate surroundings) Did you experience their presence as disturbing?	No or Yes because ... (Open question)
4. Acceptability	Please indicate what is most applicable: the technical quality of the video was ...	3 options: a) acceptable in every context, b) acceptable but only in the context in which I watched it or c) not acceptable

Table 7.4: The paper diary questions that were used to evaluate the video immediately after the playback in the living laboratory experiment, together with a reference to these questions and the possible answers

we also wanted to collect additional (contextual) information, for which the diary method is more suitable.

7.2.3 Sample description

Previous research has already indicated that the appreciation of and interest in the offered content possibly has a major impact on users' QoE [2–4]. Moreover, it has been argued that previous experiences and user-related characteristics should also be taken into account. Therefore, a specific group of users was targeted in this experiment. 30 test subjects were recruited by an experienced panel manager from iMinds-iLab.o (a research division with a strong expertise in living lab research and panel management). The recruited test subjects were meeting the three main

selection criteria: 1) being a smartphone user, 2) having watched mobile video at least once in the preceding month and 3) having indicated to have an interest in the content category used in this study (movies / movie trailers). Since the idea of a living lab implies staying close to the realistic situation, these criteria were laid down in order to reflect the natural viewing conditions and behaviour of the users as much as possible. In total, 29 people (24% female and 76% male) between 20 and 61 years old participated in the study. (The mean age is 33.1 with a standard deviation of 10.0) One test subject, who had agreed to participate, dropped out just before the actual test period. Due to time constraints, this test subject was not replaced. Every test subject received a gift voucher of 10 Euro.

In total, the data gathering phase took just over three months since the five available devices rotated among the test subjects. The data obtained via the user study were assembled and integrated into one data file containing the subjective evaluations collected through the questionnaire on the device, the paper diary entries for every question, and the logged technical data. Sessions in which video watching was not possible due to the lack of a data connection, had to be removed. Moreover, two additional sessions in which video watching was possible were removed (one outlier with an erroneous value, and one sample in which the user's ratings were missing). After excluding these sessions, 753 data samples were obtained, providing the data to analyse the viewing behaviour of the user, and to develop a model for the subjective evaluation of video quality in a mobile context.

7.3 Results

7.3.1 Viewing behaviour and subjective evaluations

In terms of physical context of the test subjects, we found that most of the videos were watched at home (82.7%) and at work (9.7%). Only 5.2% was watched during travelling. 2.4% was watched somewhere else (including, e.g., at the house of a friend or relative, in a café, or in a museum). Although one might expect that more videos would be watched during travelling, this was not the case in this study. In fact only 8 of the 29 test subjects (i.e., 27% of the test subjects) watched videos during travelling. Moreover, previous research on mobile TV points to the same observation: e.g., in [5], the results from a living lab study on mobile TV showed that most viewing occurred at home. Given the small number of samples in which movement during video playback was recorded, differences in the subjective evaluations of the quality aspects and the QoE could not be detected for different mobility states.

In terms of the acceptability of the video quality, no significant differences were found according to the physical context of the test subjects. The reason for this might be that the large majority of the videos (82.7%) were watched at home.

The answers on the question regarding the acceptability of the quality were equally distributed. 33% of the videos were evaluated as "acceptable in any context"; 33% was evaluated as "acceptable but only in the context in which I watched it"; and the remaining 34% was evaluated as "not acceptable".

Figure 7.3 shows the types of data network that were used to transfer the videos to the mobile device according to the physical location of the test subject. If inter-system handovers occurred during the video transmission, the connection type that was responsible for the majority of the video transfer is considered in Figure 7.3. 7% of the videos is transmitted on a GPRS network. Only 1% of the videos is using a UMTS connection. An EDGE connection was not used in this experiment. The most used connection type (51% of the videos) is the HSPA network, followed by the WiFi connection (41%). As shown in Figure 7.3, the type of data network is closely related to the physical context of the test subject. E.g., WiFi is almost exclusively used at home.

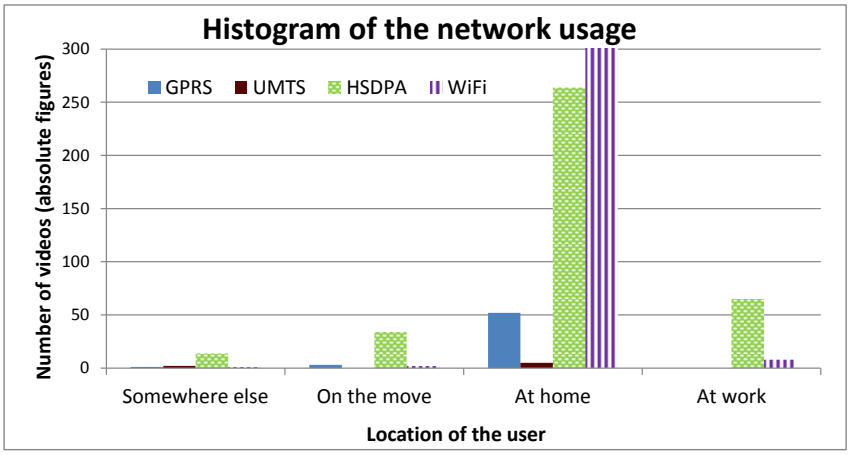


Figure 7.3: Type of data network that was used during the living lab experiment according to the location of the test subject

Time wise, Figure 7.4 shows that the evening (from 18.00 till 24.00 o'clock) was the most popular watching time, followed by the afternoon. This is the case both on week days and on weekend days. In absolute numbers, most videos were watched during the week (72.8%), which makes sense since every test subject had one week to finish the test so only two weekend days, but five week days were included in the test period. So test subjects were about equally active during weekend days as during the week days.

In 61.4% of the cases, no other people were in the immediate surroundings of the test subject (radius of approximately 5 meter) during the video watching.

22.8% of the videos were watched by the test subjects in presence of one other person. In the majority of the viewing sessions in which other people were in the surroundings of the test subject, the presence of these people was not experienced as disturbing (89.8%). In the remaining 10.2%, the talking of the others and noise made by them or coming from other sources (such as the TV) is often mentioned as disturbing factor. However, there is no significant influence of the number of people around while watching (as a variable of the 'social context' of the user) on the overall experience rating.

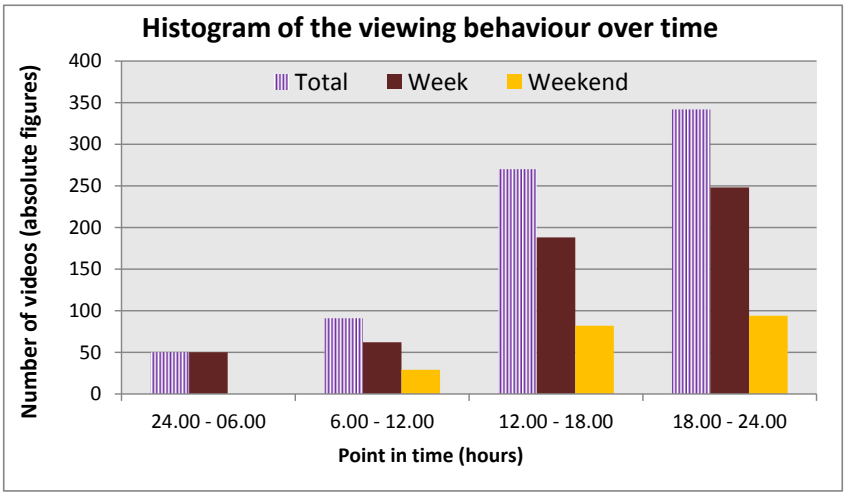


Figure 7.4: Viewing behaviour during the living lab experiment in terms of time

Figure 7.5 compares the MOS (Mean Opinion Score), i.e., the arithmetic mean of all the individual quality ratings, for the four technical combinations. Although individual ratings are ranging from very negative to very positive (as illustrated in Table 7.5 and 7.7 for the loading time and distortion), the mean values of the subjective evaluations are all quite positive and range between 2.8 and 4.1. As explained in Section 3.3.1, the assumptions of parametric hypothesis tests, such as ANOVA or T-tests, are not always fulfilled because of the discrete nature of the rating mechanism. Therefore the subjective evaluations of this experiment are also analysed using non-parametric hypothesis tests.

The subjective evaluations regarding the quality aspects of the video were compared for the four technical combinations (transport protocol and source quality) using the Wilcoxon rank sum test. In this analysis, the different technical combinations are the grouping variable (independent variable) and the subjective evaluations are the dependent variables. Significant differences ($p < .05$) were identified

for the evaluations of the technical quality, distortion, fluidity, and overall experience.

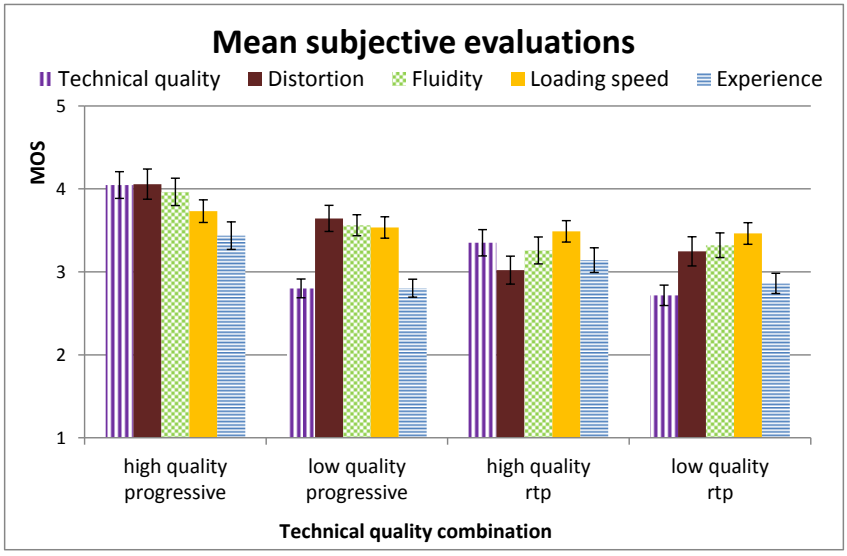


Figure 7.5: Mean subjective evaluations regarding the video quality aspects according to the four technical quality combinations

The *perceived technical quality* of the combination "high-quality video source - progressive downloading" is perceived as significantly better than that of the other combinations of video quality and transmission protocol (Figure 7.5). The technical quality of the high-quality RTP videos is evaluated as the second best option and is significantly better than the two combinations with a low-quality video source. The technical quality of the low-quality RTP videos received the lowest evaluation (Mean = 2.72; Standard Deviation = .96).

In terms of the *perceived distortion* (Figure 7.5), the differences between the high-quality progressive downloading videos and videos streamed via RTP (both low and high quality) are significant. High-quality video sessions using progressive downloading received the best evaluation regarding the perceived distortion and the difference in MOS with videos streamed via RTP is approximately 1 unit. The difference in perceived distortion between the low-quality videos transmitted via progressive downloading and the videos streamed via RTP is also statistically significant (0.62 and 0.39 on the MOS for respectively high and low-quality RTP videos). This subjectively-observed difference can be explained by the characteristics of the transmission protocol: (multiple) packet loss may induce audiovisual

distortions for video that is streamed using RTP, whereas progressive downloading based on TCP relies on retransmissions in case of packet loss.

In terms of *perceived fluidity* (Figure 7.5), the high-quality progressive downloading videos were perceived as more fluent than the streamed videos. Although the progressive downloading videos may introduce playback interruptions due to rebufferings, many of these video sessions in the experiment suffered only from a small number of short rebufferings, which were tolerated by the users. Or in the case of a fast network connection, no rebufferings at all were required.

Regarding the *perceived loading speed*, no significant difference was noticed for the various combinations of video quality and transport protocol.

The Wilcoxon rank sum test comparing the evaluations for the *overall perceived experience*, which were given in the paper diary, yields similar results for the different quality / protocol combinations: the high-quality progressive downloading videos result in a significantly higher QoE than the other combinations. The high-quality RTP videos provide test subjects the second best QoE and were evaluated significantly better than both low-quality combinations. Furthermore, the subjective evaluations showed that the overall experience of the test subjects was the worst in the case of low-quality RTP. This negative experience is in accordance with the poor evaluation of the technical parameters of the low-quality RTP videos.

7.3.2 Qualitative analysis

As the result of a qualitative analysis of the user feedback obtained via the diaries, Figure 7.6 shows the number of comments in three categories (positive aspects, negative aspects, and things that could be changed to enable a better experience) for the four video combinations.

Only for the first category of videos in Figure 7.6 (high quality - progressive), the number of positive aspects that were mentioned, supersedes the number of negative aspects and proposed changes (122 positive comments, 106 negative comments, and 49 proposed changes). Most negative feedback is given for the high-quality videos streamed using RTP (185 entries), for which the fluidity and perceived distortion was rated lowest (see Figure 7.5).

The open questions were included in the diary since it is not always clear on which specific aspects user ratings are based. Moreover, the use of numerical expressions of perceived quality is always problematic in a way since these ratings provide little insight in what this really implies from a user point of view. The answers on the open questions contain valuable information on the individual video watching sessions. First of all, they illustrate that the test subjects are precise and detailed and performed the test in a rigorous way, e.g., they make clear distinctions between different technical artefacts in their verbal evaluations. Additionally, the answers revealed that other, non-technical aspects are also considered

by test subjects when asked to reflect on positive and negative aspects of the viewing experience. Examples are issues related to the content itself (e.g., good acting, presence of a specific actor, story, emotional impact of the content, associations, ...), the sound (e.g., compelling music, aggressive sound, ...) , the colours (e.g., too bright or too dark, unnatural, ...), etc. Although the technical quality may be negatively perceived, it does not automatically result in a negative viewing experience: the experience can still be rather positive because, e.g., the user liked the music, the story, or a specific actor in the trailer. Qualitative user feedback can help to understand how the different combinations were evaluated and why one technical quality condition was preferred over another.

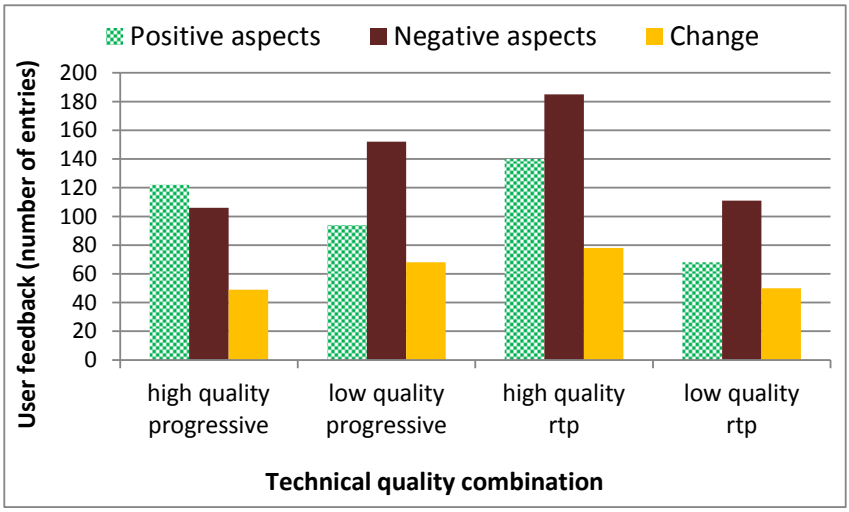


Figure 7.6: Overview of the number of qualitative user comments according to the four quality combinations

7.3.3 Modelling the subjective quality evaluations

In this section, the subjectively-perceived quality of the video sessions is further investigated in order to model the subjective evaluations based on the measured technical parameters.

7.3.3.1 Statistics used for the modelling

An important aspect during the selection of the most appropriate statistical technique is the type of data that has to be analysed. Although the answers on the multiple choice questions consist of a verbal description and a corresponding number,

it is best to consider these ratings as ordinal numbers. This means that it is possible to rank the values, but the real distance between categories is unknown. E.g., the difference between 'excellent' and 'good' may not be treated the same as the difference between 'good' and 'fair'.

Given the ordinal nature of the subjective ratings regarding the technical aspects, traditional statistical techniques, such as linear least-squares regression, are less suitable for investigating the effect of objective parameters on the rating behaviour of the users. One candidate technique to analyse the subjective ratings is ordinal logistic regression. Ordinal logistic regression is an extension of binary logistic regression (which is a model used for prediction of the probability of occurrence of an event by fitting data to a logistic function [6]). Ordinal regression modifies the binary logistic regression model to incorporate the ordinal nature of a dependent variable by defining the probabilities differently. Instead of considering the probability of an individual event, this technique considers the probability of that event and all events that are ordered before it [7].

However, one of the assumptions underlying ordinal logistic regression is that the relationship between each pair of outcome groups is the same. In other words, ordinal logistic regression assumes that the coefficients that describe the relationship between, say, the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories, etc. This is called the proportional odds assumption or the parallel regression assumption [7]. However, this test of parallel lines showed that this assumption was not valid for the obtained subjective evaluations. Therefore, different models have to be defined to describe the relationship between each pair of possible ratings by multinomial logistic regression. *Multinomial logistic regression* is also a generalization of binary logistic regression and allows more than two discrete outcomes. This regression model is used to predict the probabilities of different possible outcomes of a dependent variable (in our case the subjective rating), given a set of independent variables which may be real-valued, binary-valued, categorical-valued, etc. (in our case the objective parameters) [8]. The result of multinomial logistic regression is a comparison of the probability of a specific event against the probability of a reference event. In Section 7.3.3.2 and 7.3.3.3, multinomial logistic regression was used to model respectively the subjectively-perceived loading speed based on the measured objective loading time and the subjectively-perceived distortions based on the measured objective packet-loss rate during video playback.

The subjective experience of the user, which is assessed through the questionnaire, can be influenced by multiple objective parameters including the quality of the video, the transmission protocol, network parameters (network type and RSSI), packet-loss rate, loading time, handovers, etc. Because of these different influencing objective parameters, the QoE is modelled in Section 7.3.3.4 via a *decision*

tree, a classification technique that uses a tree-like graph or model of decisions and their possible consequences [9]. This decision support tool is in some cases preferred over other non-parametric techniques because of the readability of their learned hypotheses and the efficiency of training and evaluation.

7.3.3.2 Modelling the subjectively-perceived loading speed

One of the quality aspects that the test subjects could evaluate was the loading speed of the video. Table 7.5 shows the rating options for evaluating the perceived loading speed, the mean of the measured loading time corresponding to each option for the subjective evaluation (i.e., the mean loading time of the videos that received a specific rating), the number of video sessions that received a specific rating, and the fraction of the video sessions that received a specific rating. The loading time is measured as the time period between selecting a video and the moment when the video starts playing. The results indicate that the loading speed of the majority of the video sessions (62.4%) is evaluated as 'good' or even 'excellent'. Conversely, for a considerable part of the video sessions (15.4%), the subjectively-perceived loading speed is 'poor' or 'bad'.

Therefore, the influence of the measured objective loading time on the subjective evaluation of the perceived loading speed is investigated. Besides the loading time, the duration of the video might also influence the subjective evaluation of the loading speed. But since all videos of the experiment had approximately the same duration, this parameter is not included in the analysis.

Evaluation of the loading speed	Mean loading time (s)	Number of sessions	Fraction of the sessions
1 = Bad	29.3	59	7.9%
2 = Poor	18.7	56	7.5%
3 = Fair	5.7	167	22.2%
4 = Good	3.5	344	45.8%
5 = Excellent	2.9	125	16.6%
Total	7.1	751	100%

Table 7.5: Subjective evaluations and mean objective measurement of the loading time

For the multinomial logistic regression analysis, the subjective evaluation of the loading speed was selected as dependent, the measured objective loading time is an independent (covariate), and the reference event was the evaluation of the loading speed as 'fair'. So for each rating option, the regression model provides a function for the ratio of the probability of obtaining that specific rating, e.g., P(excellent) and the probability of obtaining the reference rating P(fair), in terms

of the measured loading time, i.e., LT. Table 7.6 lists the results of this multinomial logistic regression analysis: the probability ratios are exponential functions in terms of the measured loading time (in seconds). The likelihood ratio χ^2 of 164.7 with a p-value < 0.0001 and 4 degrees of freedom tells us that our model as a whole fits significantly better than a model without the loading time as predictor. (The χ^2 statistic is the difference in 2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model [8].)

Probability Ratio	Estimated Function
$\frac{P(Bad)}{P(Fair)}$	$exp(-1.800 + 0.068\ LT)$
$\frac{P(Poor)}{P(Fair)}$	$exp(-1.652 + 0.060\ LT)$
$\frac{P(Fair)}{P(Fair)}$	1
$\frac{P(Good)}{P(Fair)}$	$exp(1.075 - 0.081\ LT)$
$\frac{P(Excellent)}{P(Fair)}$	$exp(0.261 - 0.143\ LT)$

Table 7.6: The results of the multinomial logistic regression analysis with the subjective evaluation of the loading speed as dependent and the measured objective loading time as a covariate (LT = loading time)

Figure 7.7 visualizes these probability ratios for a measured loading time between 0 and 40 seconds. The graph shows that for short loading times (less than 10 seconds), a high probability exists that users will evaluate the loading speed as 'good' or 'excellent'. Given the large fraction of video sessions evaluated as 'good' (45.8% in Table 7.5), the probability of obtaining 'good' as subjective evaluation is higher than the probability of obtaining 'excellent'. If the measured loading time is more than 13 seconds, users are more willing to evaluate the loading speed as 'fair' than to rate it as 'good'. For short loading times, users are not inclined to give low evaluations like 'bad' or 'poor'. However after a loading time of approximately 27 seconds, ratings with the label 'bad' or 'poor' are more likely than the reference rating, i.e., 'fair'. And for instance after 40 seconds of loading time, it is 2.5 times more likely that users perceive the loading speed as 'bad' than that users perceive it as 'fair' (Figure 7.7).

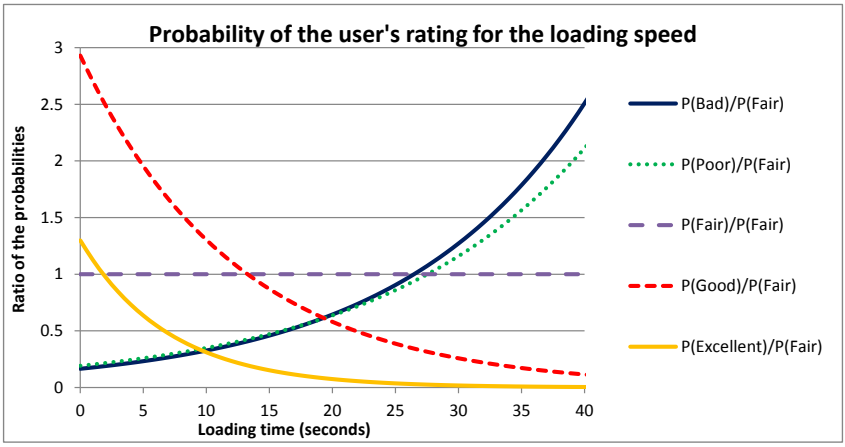


Figure 7.7: The probability ratios of the ratings options for the perceived loading speed

7.3.3.3 Modelling the subjectively-perceived distortion

In contrast to progressive download, which relies on packet retransmissions in case of packet loss, video streaming via RTP might suffer from audiovisual distortions if packets are lost during transmission. Therefore, the influence of packet loss on the subjectively-perceived distortion during mobile video watching was investigated for the video sessions which are streamed via RTP. Table 7.7 shows the rating options for evaluating the perceived distorting during video watching, the mean of the measured packet-loss rate corresponding to the subjective evaluation (i.e., the mean packet-loss rate of the videos that received a specific rating), the number of video sessions that received a specific rating, and the fraction of the video sessions that received a specific rating. This analysis was based on the data samples obtained for the mobile video sessions using RTP (high- and low-resolution videos).

Table 7.7 shows that sessions which received a positive evaluation regarding the perceived distortion ('not noticeable' or 'noticeable, not annoying') are characterized by a low packet-loss rate (mean values of 0.8% and 0.4%). In contrast, low ratings for the perceived distortion ('noticeable, annoying' or 'noticeable, very annoying') are typically due to high packet-loss rates (mean values of respectively 18.9% and 32.5%). Therefore, the influence of this packet-loss rate on the subjectively-perceived distortion during mobile video watching is further investigated.

As already indicated in Chapter 5, the perceived distortion due to packet loss depends on various technical parameters of the video, such as the codec. So if e.g., another video codec is used, the results can slightly change and the analysis should be repeated. Moreover, different individual packet losses can have a different impact on the perceived distortion due to patterns of subsequent packet losses (bursts)

or the type of frame in which packet loss occurs (I, P, or B-frame). However for this analysis, individual packet losses are not investigated in detail, but the effect of a substantial packet loss rate as a whole is investigated.

Evaluation of the distortion	Mean packet-loss rate	Number of sessions	Fraction of the sessions
1 = Noticeable, very annoying	32.5%	71	19.1%
2 = Noticeable, annoying	18.9%	67	18.0%
3 = Noticeable, slightly annoying	3.1%	78	21.0%
4 = Noticeable, not annoying	0.4%	68	18.3%
5 = Not noticeable	0.8%	88	23.7%
Total	10.5%	372	100%

Table 7.7: Subjective evaluations of the distortion and mean objective measurement of the packet-loss rate

For the same reason as in the analysis of the loading speed, a multinomial logistic regression analysis was performed to estimate the probability of obtaining a specific rating as a function of the packet-loss rate. For this analysis, the subjective evaluation of the perceived distortion was selected as dependent, the measured objective packet-loss rate is an independent (covariate), and the reference event was the evaluation of the distortion as 'noticeable, slightly annoying'. For each rating option, Table 7.8 lists the ratio of the probability of obtaining that specific rating, e.g., P(not noticeable), and the probability of obtaining the reference rating, P(noticeable, slightly annoying), in terms of the measured packet-loss rate, i.e., PL. The likelihood ratio χ^2 of 149.3 with a p-value < 0.0001 and 4 degrees of freedom tells us that our model as a whole fits significantly better than a model without the packet-loss rate as predictor.

Figure 7.8 visualizes the probability ratios of Table 7.8 for a packet-loss rate ranging from 0% to 40% (using a logarithmic scale). Video sessions with a limited packet-loss rate have a higher probability to obtain a positive rating regarding the perceived distortion ('not noticeable' or 'noticeable, not annoying') than to receive the reference rating (i.e., 'noticeable, slightly annoying'). Around a packet-loss rate of 0.25%, the probability of a positive evaluation starts decreasing. When more than 2.6% of the packets are lost during transmission, the probability that users are slightly annoyed by distortions is higher than the probability that users do not notice these distortions (solid decreasing line versus dashed horizontal line in Figure 7.8). If the packet-loss rate during video watching is higher than 30%, the probability of receiving a positive evaluation from the user is very small (less than 5% of the probability of receiving the reference rating).

Probability Ratio	Estimated Function
$\frac{P(Noticeable,\ very\ annoying)}{P(Noticeable,\ slightly\ annoying)}$	$exp(-0.903 + 0.072\ PL)$
$\frac{P(Noticeable,\ annoying)}{P(Noticeable,\ slightly\ annoying)}$	$exp(-0.609 + 0.058\ PL)$
$\frac{P(Noticeable,\ slightly\ annoying)}{P(Noticeable,\ slightly\ annoying)}$	1
$\frac{P(Noticeable,\ not\ annoying)}{P(Noticeable,\ slightly\ annoying)}$	$exp(0.147 - 0.287\ PL)$
$\frac{P(Not\ noticeable)}{P(Noticeable,\ slightly\ annoying)}$	$exp(0.302 - 0.115\ PL)$

Table 7.8: The results of the multinomial logistic regression analysis with the subjective evaluation of the distortion as dependent and the measured objective packet-loss rate as a covariate

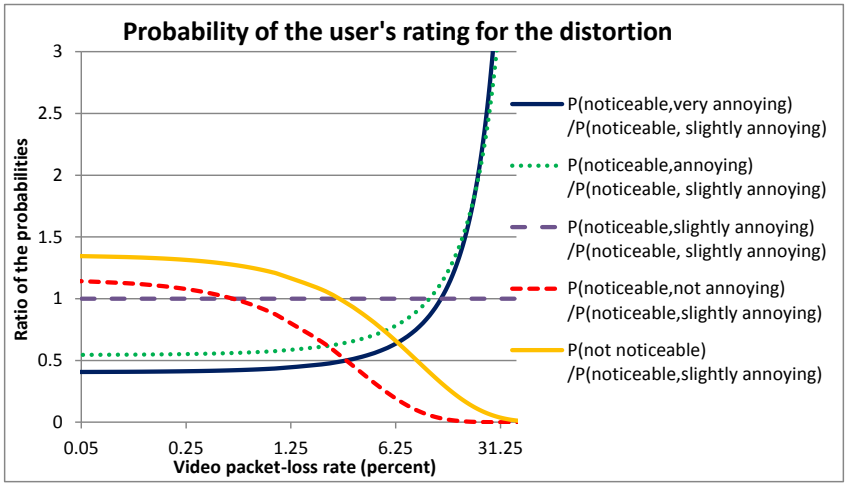


Figure 7.8: The probability ratios of the ratings options for the perceived distortion

Negative evaluations of the perceived distortion are less likely than the reference rating for low values of the packet-loss rate. On the other hand, the rating options 'noticeable, annoying' and 'noticeable, very annoying' are more likely than the reference option 'noticeable, slightly annoying' as soon as the packet-loss rate is higher than respectively 10.5% and 12.5%.

7.3.3.4 Modelling the subjectively-perceived experience

In Section 7.3.3.3 the influence of packet-loss on the perceived distortion was investigated and in Section 7.3.3.2 the subjectively-observed loading speed is mod-

elled based on only the measured loading time. In this section, multiple objective parameters are considered to model the subjectively-perceived experience of the user during video watching via a decision tree.

Each of the gathered objective parameters, as listed in Table 7.2, is a potential parameter on which to base the decisions that have to be made in the decision tree. The minimum set of technical parameters required to model the QoE was determined through a statistical analysis. The resulting parameters regarding the video session that are used as input for the decision tree are: the transport protocol, the quality of the video source, the types of data network that were used to transmit the video, the number of handovers during transmission, and the percentage of the video that was actually watched by the user. Because of the mutual correlation[1] of the technical parameters, additional parameters, such as packet loss, jitter, mobility of the user, and signal strength of the network, have no additional information value and do not further improve the classification model; i.e., the inclusion of these parameters in the decision tree did not lead to a further decrease of the misclassification rate.

Figure 7.9 shows the visualization of the decision tree, which can be used to predict the user's experience based on the technical parameters of the video session. To avoid overfitting and limit the complexity of the model, the decision tree was pruned until a deviance[2] of 0.008 was obtained. The starting point of this decision tree is the root, situated at the top of the figure.

At the first fork, a decision is made based on the quality of the video source: if the video source has a low quality, the left branch is chosen; the right branch is followed, in case of a high-quality video source. These low-quality videos typically induce a poor to fair QoE: the estimated experience ratings for low-quality videos are ranging from 1 to 3 points in the decision tree. To predict the QoE while watching low-quality video sources, the type of data network used for transmitting the video is important. If the video is mainly transmitted over a GPRS connection (more than 90% of the video is transmitted over a GPRS connection, so less than 10% is transmitted over a faster data connection), the decision tree predicts a QoE value of 1 point. The reason for this bad experience may be the combination of a low-quality video source and a slow data connection type (GPRS). This slow data connection may introduce interruptions during video playback (i.e., long rebuffering times if progressive download is used or audiovisual distortions if the video is streamed). If the low-quality video is (partly) transmitted over a faster data network, an additional criterion is investigated to estimate the user's experience.

[1] If two parameters are mutually correlated, they are related to each other somehow, but not necessarily by cause and effect.

[2] Prune on deviance is a measure that defines a stopping rule in the pruning process based on maximum-likelihood principles.

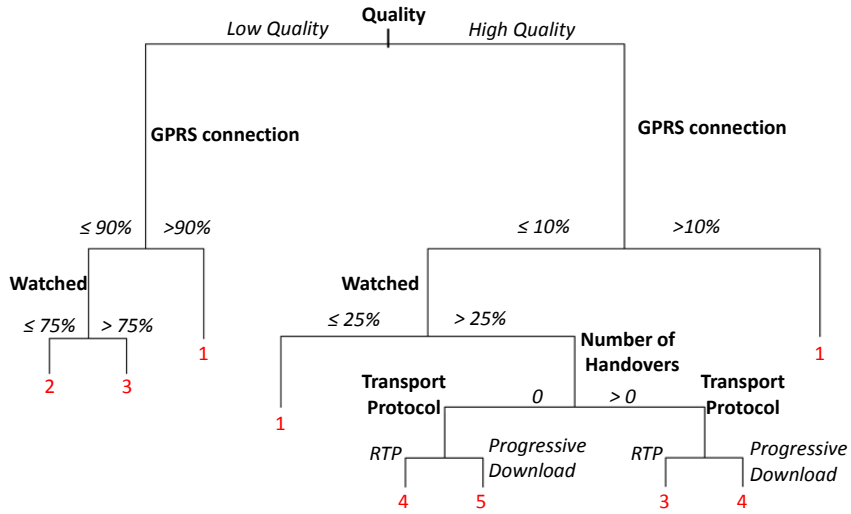


Figure 7.9: Decision tree modelling the QoE during video watching on a mobile device, based on the watching behaviour and the technical parameters of the video and network

The next branch is based on the part of the video that was actually watched by the user. On the one hand, if the user watched the video (almost) completely (more than 75%), this may suggest that no network problems occurred during the video playback and the user experienced the video session as acceptable. In this situation, the user will rate his/her experience with 3 points, according to the decision tree. On the other hand, a video session that is stopped early by the user (less than 75% of the video is watched) can indicate a bad experience due to an unacceptable audiovisual quality or a disinterest of the user for the content of the video. Alternatively, the video stoppage might be due to a network problem (e.g., a network disconnection), which also induces a bad experience for the user. Therefore the decision tree provides a prediction of 2 points for the QoE in this condition.

In case of a high-quality video source, the right half of the decision tree is used to classify the QoE. A GPRS connection is basically too slow to transmit the high-quality video, thereby introducing distortions or interruptions during video playback. So, if a GPRS network is used for (more than 10% of) the transmission of the high-quality video, the QoE is estimated to be bad, which is indicated by a prediction of 1 point for the experience value. The other branch represents the situation of transmitting a high-quality video source over a fast data network (no or limited GPRS is used). In this situation, the QoE is predicted based on the fraction of the video that is actually watched by the user. Videos might be stopped by the user because of a bad experience (due to a low audiovisual quality or uninteresting

content). In addition, video stoppages / interruptions due to network problems can be the cause of a bad QoE. As a result, these early video stoppages lead to a prediction value of 1 point for the QoE.

High-quality videos that are transmitted over a fast network without an early video stoppage (more than 25% of the video is watched) indicate a good QoE. In this situation, the experience value is predicted based on the used transmission protocol and the number of handovers. High-quality video sources streamed over the cellular data network may via RTP suffer from packet loss and jitter, inducing video distortions. Progressive download can resolve these packet losses and avoid distortions by using retransmissions but may thereby introduce extra rebuffering times. The subjective evaluations of the test subjects learned that in this experiment users have a better experience with videos transmitted using progressive download, than with streamed videos (using RTP). This finding confirms the results of a study on the impact of the underlying transport protocol on the QoE for streaming media services [10]. Therefore, the experience of video sessions using progressive download is predicted to be (1 point) better than the experience of video sessions based on streaming.

During a handover, an ongoing data session is transferred from one channel or cell to another one. Since this process may introduce an interruption of the data transmission, it can have an influence on the QoE. As a result, the decision tree predicts a lower experience value for video sessions with handovers than for sessions without handovers.

This decision tree can be used to predict and improve the user's QoE of a mobile video session. E.g., if a fast cellular data network is available for the transmission of half of the video, (i.e., 50% of the video is transmitted over a GPRS network and 50% is transmitted over a faster network such as UMTS or HSPA), a low-quality video source might be preferred above a high-quality video source. In this situation, the decision tree predicts an experience value of 1 point for the high-quality video source, whereas the low-quality video source receives 2 or 3 points depending on the fraction of the video that is actually watched. In contrast, if a fast cellular data network is available for the complete transmission of the video (and video playback is not stopped early), a high-quality video source might result in a better QoE, as indicated by the decision tree.

To train the decision tree, 90% of the data samples were used as training set (i.e., 616 randomly selected samples); the remaining 10% (68 samples) constitutes the test set and was utilized for the validation of the decision tree. (Since the test subjects specified their experience during video watching not for all 753 video sessions, only 684 samples were available to train and validate the decision tree). After training the decision tree, the samples in the test set were classified according to the obtained decision model and the misclassification rate was calculated. This procedure was repeated 20 times to eliminate influences of the random data

partitioning in training and test set. The averages (arithmetic mean) of the misclassification rates, obtained during these 20 iterations, are indicated in Table 7.9. For almost half of the test samples (46.2%), the decision tree is able to correctly predict the QoE rating provided by the user based on the watching behaviour, transport protocol, and the technical parameters of the video and the network. Moreover the QoE of 81.5% of the video sessions is classified correct (46.2%) or within an acceptable error margin of 1 point (35.3%). The ratio of severe misclassifications (i.e., 3 or 4 points deviance between the predicted experience and the actual experience) is limited to 2.3% of the samples, which proofs the usefulness of the decision tree.

Classification	Ratio of test samples
Correct classification	46.2%
1 point misclassification	35.3%
2 points misclassification	16.2%
3 points misclassification	1.6%
4 points misclassification	0.7%

Table 7.9: Misclassification rate of the decision tree

7.4 Conclusions

In this exploratory study drawing on the evaluation of objective and subjective QoE aspects by a user panel, we investigated Quality of Experience (QoE) related to mobile video watching in a semi-living lab environment. 28 video trailers were watched by the test subjects in random combinations of two video resolutions (high and low) and two data transfer protocols for video (RTP and progressive download using TCP/HTTP). The participants were able to watch the videos when they wanted, where they wanted and user evaluations were gathered by means of questionnaires on the device, complemented with traditional pen and paper diaries. The results illustrate that most videos were watched at home and in the afternoon and evening. In most cases, no other people were around during the watching session. The presence of other people did not have a significant influence on the overall experience rating and was in 90% of the cases not perceived as a disturbing factor.

A statistical analysis compared the subjective quality ratings for the four technical quality combinations. Both the qualitative and quantitative feedback showed that the high-quality progressively downloaded videos yield a significantly better experience than the streamed videos in terms of perceived technical quality, distortion, fluidity, and overall experience. The technical quality of the low-quality

video sources using RTP was evaluated as the worst. Analysis of the qualitative user feedback could help to understand which aspects influenced the overall QoE in a positive and negative way in the four technical quality combinations.

The influence of the measured loading time on the subjective evaluations of the loading speed was evaluated via a multinomial logistic regression analysis. The resulting model showed that when the loading time increases from 10 to 30 seconds, the subjective evaluations of the loading speed gradually evolve from mainly positive to mainly negative.

For video sessions using RTP, we investigated the subjectively-perceived distortion during mobile video watching as a function of the video packet-loss rate. The probability of receiving a positive rating is rapidly decreasing as more packet loss occurs during video watching (from a packet-loss rate of around 0.25%) and video sessions with a packet-loss rate of more than 10% are in general evaluated as 'annoying' or even 'very annoying'.

Finally, this measurement study resulted in a decision tree for quantifying QoE during mobile video watching based on objective parameters such as network and video quality. This model provides application developers and service providers a tool that clarifies which and how technical parameters influence the QoE and how the parameters have to be adapted to optimize the QoE. E.g., if no fast cellular data network is available (only a GPRS network is available), the model predicts a bad QoE for high-quality video sources due to distortions or a large amount of rebufferings during video playback. In this case, low-quality video sources might result in a better QoE. If a fast data network is available, a high-quality video transmitted using progressive download provides an optimal QoE.

The presented study can be seen as an example of QoE research in a real-life, semi-living lab setting. Given the increased emphasis on contextual variables and subjective, user-related characteristics of QoE, new context-aware tools and measurement approaches should be explored to take these dimensions into account. Whereas research in controlled settings is very valuable to assess the influence of particular, isolated parameters, research in more natural and ecologically valid settings might help to better understand the interplay between different parameters and their relative influence on the overall QoE.

References

[1] *P.911: Subjective audiovisual quality assessment methods for multimedia applications*. Technical report, ITU-T, International Telecommunication Union, 1998. Online available at `http://www.itu.int/rec/T-REC-P.911-199812-I/en`.

[2] P. Kortum and M. Sullivan. *The Effect of Content Desirability on Subjective Video Quality Ratings*. Human Factors: The Journal of the Human Factors and Ergonomics Society, 52(1):105–118, 2010.

[3] P. Kortum and M. Sullivan. *Content is King: The Effect of Content on the Perception of Video Quality*. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 48(16):1910–1914, 2004.

[4] S. H. Jumisko, V. P. Ilvonen, and K. A. Vaananen-Vainio-Mattila. *Effect of TV content in subjective assessment of video quality on mobile devices*. In R. Creutzburg and J. H. Takala, editors, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, volume 5684 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 243–254, March 2005.

[5] D. Schuurman, K. De Moor, L. De Marez, and T. Evens. *A Living Lab research approach for mobile TV*. Telematics and Informatics, 28(4):271 – 282, 2011. Television in a digital era - Usage and policy issues.

[6] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. McGraw-Hill, fifth edition, 2005.

[7] *SPSS Data Analysis Examples Ordinal Logistic Regression*. Technical report, Academic Technology Services, Statistical Consulting Group, University of California (UCLA), 2012. Online available at `http://www.ats.ucla.edu/stat/spss/dae/ologit.htm`.

[8] T. F. Liao. *Interpreting Probability Models, Logit, Probit, and Other Generalized Linear Models*. SAGE Publications, Newbury Park, first edition, 1994.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, first edition, January 1984.

[10] T. Hoßfeld, R. Schatz, T. Zinner, M. Seufert, and P. Tran-Gia. *Transport Protocol Influences on YouTube Videostreaming QoE*. Science research report series, University of Würzburg Institute of Computer, 2011. Online available at `http://www3.informatik.uni-wuerzburg.de/TR/tr482.pdf`.

8

The influence of QoE on the rating behaviour

8.1 Introduction

Recommendation techniques share a common prerequisite: user preferences or personal tastes regarding previously consumed content are required to generate personal suggestions. These user preferences can be retrieved through implicit or explicit feedback mechanisms [1]. However, a serious problem of these feedback mechanisms is the discrepancy between the received feedback value, and the actual interests of the user. Besides the content and personal interests, user feedback on audiovisual material can be influenced by additional factors such as the picture and sound quality, fluidity of the video, loading speed, distortions, etc. Studies have already shown a strong correlation between user feedback regarding the content and subjective evaluations of the video quality [2]. So, technical parameters and audiovisual quality aspects might influence the user's explicit feedback, which is supposed to reflect his/her personal interests and to evaluate merely the content of the video. However according to our knowledge, no solution is available to correct for the influence of these technical parameters on the explicit feedback.

The efficiency of personal video suggestions generated by recommender systems is highly dependent on the quality of the obtained user feedback. This feedback has to reflect the personal interests in the content of the viewed video, to obtain accurate recommendations. Consequently, the incorrect estimation of the user's actual interests in the content might erroneous update the preferences in the

user profile thereby leading to inaccurate recommendations. To date, this issue has received very little research attention.

8.2 Test setup

8.2.1 Goals of the study

The goal of this research is to investigate the influence of measured objective parameters and audiovisual quality aspects on the experience of users and the coupled effect on the explicit feedback behaviour in a mobile context. Via a living lab experiment, technical parameters are monitored and related to the subjective experience of the end-user while watching streaming video on a mobile device in a real-life setting.

According to our knowledge, no work has been done to correct the users' explicit feedback for video content by considering the influence of a varying audiovisual quality. This experiment resulted in a feedback model for recommender systems which takes into account the technical parameters of the mobile network and the video transmission, and the effect on the user's rating behaviour. This model can be used as an additional feedback filter for video recommender systems that could help to eliminate the influences of audiovisual quality on explicit user feedback.

8.2.2 Procedure

For this experiment, the test subjects were asked to use 'PersonalTV Mobile', the mobile client application of the service discussed in Chapter 2. Via the PersonalTV Mobile application, test subjects could select, watch, and evaluate streaming videos from YouTube on a smart phone in their everyday environment, where and when they wanted (i.e., in a living lab context).

Given the importance of the content and the users' preferences for the content in this experiment, the content was not predefined (like in the experiments of Chapter 6 and 7), but test subjects could search and select the videos from YouTube's content catalog according to their preferences. This way, a much broader range of content was available than in previous experiments, enabling all test subjects to find videos that optimally suit their personal tastes. Driven by the aim of monitoring technical parameters in detail, which is not possible on the standard video clients for mobile devices, we developed our own mobile video application. Using PersonalTV Mobile instead of the standard YouTube application enables the logging of these technical parameters (Table 7.2), the same parameters that were measured in the living lab experiment of Chapter 7.

The procedure followed in this experiment consists of 3 successive phases.

8.2.2.1 Phase 1: profile building

Firstly, the test subjects were asked to watch and evaluate some YouTube videos of their own choice, on their personal computer using the desktop client of the PersonalTV application, as discussed in Chapter 2. This desktop client, which was implemented as a Facebook application for authentication reasons, has the same look and feel as its mobile counterpart. By using this desktop application, test subjects explored the PersonalTV menus and got familiar with the interface. Moreover, the PersonalTV service could build a profile for every test subject, based on the personal watching and rating behaviour. Using these user profiles, the PersonalTV service was able to avoid the cold start problem [3] and calculate personal suggestions for the mobile experiment.

8.2.2.2 Phase 2: instruction meetings

Secondly, the test subjects were divided in groups of five people and received a briefing about the experiment. Similar to the procedure described in Section 7.2.2, these instruction meetings were organized to explain the use of the application before the actual test started. For this experiment, test subjects received an HTC Android Developer Phone 1 (ADP1), since the Nexus One devices that were used in the experiments of Chapter 6 and 7 were not available during the test period. Nevertheless since these ADP1 devices are also running on the Android operating system (version 1.5), differences in the operation and usage of the phones are limited. After filling in a short general questionnaire (including, e.g., socio-demographical questions, general questions on their current use of (mobile) online video sites, attitudes, etc.), they were invited to try the PersonalTV Mobile application for the first time using the ADP1 devices.

8.2.2.3 Phase 3: mobile video watching in a living lab environment

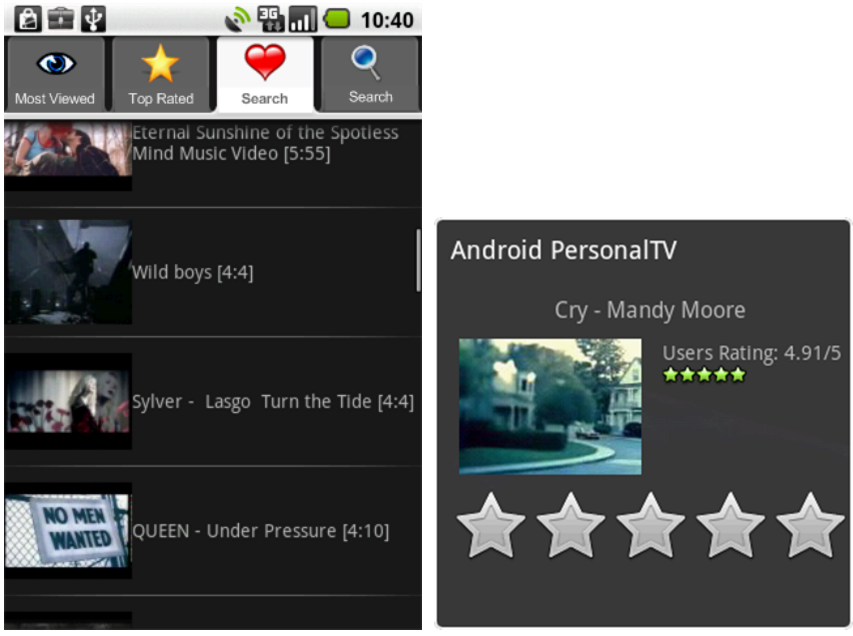
Thirdly, every test subject took an ADP1 device at home to use PersonalTV Mobile in their daily environment during the next three days. The videos, originating from YouTube, were covering a wide range of content categories such as sports, entertainment, music, comedy, technology, etc.

Figure 8.1(a) shows a screenshot of the PersonalTV Mobile application displaying the main menu, consisting of four tabs, each containing a list of videos with a thumbnail, the title, and the duration as additional information. The first two tabs offer respectively the most viewed and top rated YouTube videos within a selected period of time (today, last week, last month, or all time). The third tab presents a set of personal video suggestions, recommended by the PersonalTV service. These recommendations are based on the personal preferences of the end-user as expressed by the ratings. Ideally, this user feedback should be adjusted to compensate for the influence of objective, technical parameters of the network

and video session, as will be described in Section 8.3. However, to obtain data that allow to investigate this influence, this quality-based adjustment was not enabled during the user tests. Standard keyword-based searching among the available YouTube videos can be performed using the fourth tab, listing the results either by relevance, view count, or average rating.

For the experiment, test subjects were asked to select and watch (at least) 10 videos from their personal video suggestions generated by the application. These video suggestions might better match the personal preferences of the test subjects than random or the most-popular videos. During this video watching, the objective parameters of Table 7.2 were monitored on the mobile device.

After video watching, PersonalTV Mobile offers users the possibility to evaluate the video through a 5-point scale star-rating mechanism, similar to the rating mechanism of the desktop client of PersonalTV as illustrated in the screenshot of Figure 8.1(b).



(a) The video selection mechanism

(b) The rating mechanism that is shown after each video

Figure 8.1: Screenshots of the mobile PersonalTV application

Table 8.1 lists the characteristics of the videos, as used in the PersonalTV Mobile application. To investigate the influence of the QoE on the user's explicit feedback expressed by the star-rating mechanism, the technical parameters of the video

are chosen rather low in comparison with the parameters of Table 6.2 and 7.1. E.g., the total bandwidth requirement is 64 kbit/s which is rather low even in a mobile context [4, 5]. Moreover, all videos are transmitted using RTP to optimally study the influence of packet loss and jitter in this experiment.

Video Source			
Audio		Video	
Codec	AMR-NB	Codec	H.263 2000
Average bit rate	12 kbit/s	Average bit rate	52 kbit/s
Maximum bit rate	18 kbit/s	Maximum bit rate	230 kbit/s
Channels	1	Resolution	176 x 144
Sampling frequency	8000 Hz	Frame rate	15 fps

Table 8.1: Technical parameters of the mobile video used to investigate the influence of QoE on the rating behaviour

Finally, a small questionnaire is shown on the device, immediately after watching the video. As in the experiments described in Chapter 6 and 7, these questions pop-up on the screen after the video playback and users have to answer them before the next video can be played. Similar to the questionnaire of Table 7.3, the questions probe the users' subjective evaluations of various quality aspects of the watched video in detail. Furthermore, users were asked to evaluate the content of the video on a 5-point scale, as if the video is available in perfect quality.

8.2.3 Sample description

For this experiment, we recruited 29 test subjects using a convenience panel-sampling method. 79.3% of them were male and 20.7% was female. This majority of male test subjects is largely due to the fact that we mainly recruited test subjects at the Faculty of Engineering from Ghent University, which has a large majority of male students and researchers. The age of the test subjects ranged between 23 and 36, with a mean age of 28.0 years old (standard deviation is 3.8). The majority of the test subjects were employees (10.3% was still a student).

The experiment resulted in a total number of 392 observations, containing metadata of the selected videos (title, category, URL, and tags), contextual information (time stamp, location, and user movement), as well as objective (measured) and subjective (questionnaire) parameters. This total number of 392 observations, resulting from the 29 test subjects is considered sufficient for the central limit theorem to hold and ascertain normality of the residuals and estimated regression coefficients following from linear regression [6]. By this, the distributional as-

sumptions of the T-tests for the significance of regression coefficients are satisfied, which makes these tests meaningful.

8.3 Results

Table 8.2 lists the correlations between some important, measured objective parameters of the video session, and the user's subjective star-rating for the video. As explained earlier, recommender systems extract the user's preferences for the content from these ratings in order to generate personal suggestions for the user. As a result, these ratings should reflect the user's tastes, independent of the technical quality of the video playback. However, Table 8.2 shows a significant negative (Pearson) correlation between the technical aspects and the ratings, indicating a dependency between the objective parameters and the subjective feedback. (P-values below 0.05 indicate a significant correlation.)

	Correlation with the Star-Rating (SR)	p-value
GPRS Percentage (GP)	-0.142	0.002
mean Video Jitter (VJ)	-0.111	0.016
mean Audio Jitter (AJ)	-0.111	0.016
Video packet-Loss rate (VL)	-0.171	0.000
Audio packet-Loss rate (AL)	-0.161	0.001

Table 8.2: Correlations between the measured objective parameters of the video session and the subjective rating for the video

Traditional recommender systems, which are based on the user's rating behaviour, will perform below par if these ratings are influenced by quality aspects of the video. After all, the audiovisual quality of streaming video is mainly determined by the data connection (type) and the network conditions, which are very dependent on the (spatio-temporal) context of the end-user (and other users active on the network). As a result, a user might rate the same video differently, according to differences in the technical parameters of the video session.

As a solution to this undesired effect, the user's star ratings should be corrected and the influence of the audiovisual quality should be eliminated. This correction can be done by a linear regression model that infers the user's preferences for the content, based on the obtained star-rating and the measured objective parameters of the video session. So based on the data obtained in this experiment, a regression analysis was performed with the user's evaluation of the content as dependent variable and the measured objective parameters and the star-rating as independent variables, resulting in the following model.

$$C = 0.5094 + 0.8795\,SR + 0.0086\,AL + 0.0062\,GP + 0.0008\,VL + 0.0000\,VJ + 0.0000\,AJ \tag{8.1}$$

This regression model (8.1) expresses the user's personal preferences for the content (C) (as stated in the questionnaire) in terms of the subjective star-rating (SR), audio packet-loss rate (AL), the percentage of the video streamed over a GRPS connection (GP), video packet-loss rate (VL), mean video jitter in seconds (VJ), and mean audio jitter in seconds (AJ). This model has an R^2 value of 0.67. R^2 is the coefficient of determination and stands for the proportion of variability in the data set that is accounted for by the statistical model. Because of the mutual, positive correlation between the objective parameters, the regression model can be simplified by eliminating insignificant predictors of the user's personal preferences for the content. Using a stepwise regression analysis with bidirectional elimination, the predictor variables AL, VL, VJ, and AJ are removed from the model resulting in a more compact formula:

$$C = 0.4887 + 0.8750\,SR + 0.0091\,GP \tag{8.2}$$

The resulting model of (8.2) has still an R^2 value of 0.67 and predicts the user's personal preferences for the content (C) based on merely the subjective star-rating of the user (SR) and the percentage of the video streamed over a GRPS connection (GP).

The high (and significant) regression coefficient of SR in (8.2) indicates that the user's personal preferences for the content are mainly determined by his/her general star-rating. The (significant) regression coefficient of GP stands for the influence of the technical parameters on the user's star-rating. Videos that are streamed over a GPRS connection are typically suffering from interruptions or distortions during video playback, which influence the user's star-rating. Because of these technical difficulties during playback, the content of the videos might be underestimated by the traditional star-rating. The resulting model (8.1) tries to correct this influence by taking into account the technical aspects of the network, which are all reduced to the connection type used for the video streaming in model (8.2). So, if a video is streamed (partially) over a GPRS connection, the user's true preference for the content is estimated significantly higher than the provided star-rating. E.g., a rating of 2 stars on a video that is streamed over a GPRS connection ($GP = 100$) is estimated to correspond to a true preference for the content of 3.15.

To validate the proposed regression model, an estimation of the user's personal preferences for the content was calculated with (8.2) for each video watching session of the experiment using the cross-validation technique. Subsequently, these estimations of the user's personal preferences were compared with the user's actual evaluation of the content as expressed by the user through the questionnaire.

Table 8.3 summarizes the results of this evaluation by listing the root mean square error (RMSE) of these estimations with respect to the actual evaluations of the content, and the number of samples used for the evaluation. (Because of some missing values for the user's evaluation of the content, the model could not be evaluated on all 392 samples of the data set.)

The regression model is benchmarked against the traditional star-rating approach, which does not take the network conditions into account but estimates the user's preference for the content based on only the provided star-rating. The results of Table 8.3 demonstrate that the RMSE of the regression model is lower than the RMSE of the star-rating approach, which proves that the estimations of the regression model are a better reflection of the user's actual preferences than the traditional star-rating approach.

For a more in-depth validation, the estimations of the regression model are rounded to the nearest integer value and compared with their actual evaluations of the content. Table 8.3 lists the number of rounded estimations that equal the actual evaluations (reported as correct predictions). The incorrect estimations are classified according to their deviation from the actual evaluation (1, 2, 3 or 4-stars errors). Table 8.3 indicates that 92% ($\frac{336}{364}$) of the content evaluations are estimated correctly or with a deviation of a single star by the regression model. Only a minority of the estimations (3 for the regression model) deviate more than 2 stars from the user's actual evaluation of the content.

Again, a comparison with the traditional star-rating approach is made. As indicated in Table 8.3, fewer errors are made by the regression model than by the star-rating approach, which proves the usefulness of the regression model for estimating the user's preferences.

	Star-rating model	Regression model
Number of evaluations	364	364
RMSE	0.82	0.74
Number of correct predictions	238	249
Number of 1-star errors	89	87
Number of 2-star errors	32	25
Number of 3-star errors	5	3
Number of 4-star errors	0	0

Table 8.3: Evaluation of the regression model and the traditional star-rating mechanism

8.4 Conclusions

This chapter discussed the influence of objective, technical parameters of the video session on the user's explicit feedback, in the form of a star-rating. In this living lab experiment, test subjects were asked to select, watch, and rate streaming videos on a smartphone, while the technical parameters of the network are monitored.

Correlations between the measured objective parameters and the users' star-ratings are in line with the assumption that users tend to give a lower star-rating if the quality of the video playback is not optimal. This can have a serious impact on the accuracy of recommender systems, which assume that the star-rating reflects merely the user's preferences for the content regardless the technical parameters of the network during video playback.

Based on the obtained data samples, a regression model was proposed to correct for the influence of these technical parameters. Using a stepwise regression analysis, the connection type used during video transmission turned out to be an important factor for quantifying this influence. Validation of the model showed that the proposed regression model generates more accurate estimations of the user's actual preferences for the content, than the traditional star-rating mechanism.

This model can be used by video recommender systems to improve the accuracy of user feedback by eliminating the influence of a varying audiovisual quality induced by changing network parameters. As the user feedback better reflects the personal preferences for the content, personal recommendations become more accurate.

Future research can comprise the refinement of the model by including additional influencing factors such as the activity, location, and expectations of the end-user. Moreover, incorporating the proposed model in a recommender system followed by evaluating the efficiency of the recommendations is an interesting possibility for future work in this research domain. This way, traditional recommendations based on the star-rating mechanism can be compared with recommendations based on the proposed regression model.

References

[1] D. Oard and J. Kim. *Implicit Feedback for Recommender Systems*. In Proceedings of the AAAI Workshop on Recommender Systems, pages 81–83, 1998.

[2] P. Kortum and M. Sullivan. *The Effect of Content Desirability on Subjective Video Quality Ratings*. Human Factors: The Journal of the Human Factors and Ergonomics Society, 52(1):105–118, 2010.

[3] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. *Methods and metrics for cold-start recommendations*. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '02, pages 253–260, New York, NY, USA, 2002. ACM.

[4] S. Jumisko-Pyykkö and J. Häkkinen. *Evaluation of subjective video quality of mobile devices*. In Proceedings of the 13th annual ACM international conference on Multimedia, MULTIMEDIA '05, pages 535–538, New York, NY, USA, 2005. ACM.

[5] H. Knoche, J. D. McCarthy, and M. A. Sasse. *Can small be beautiful?: assessing image resolution requirements for mobile TV*. In Proceedings of the 13th annual ACM international conference on Multimedia, MULTIMEDIA '05, pages 829–838, New York, NY, USA, 2005. ACM.

[6] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. McGraw-Hill, fifth edition, 2005.

9 Conclusions and future research

This chapter summarizes the conclusions obtained in the research chapters 2 to 4 and 6 to 8. Part of this chapter is also dedicated to opportunities for future research in the domain of recommender systems and QoE.

9.1 Conclusions

In this dissertation, two topics are investigated for the purpose of optimizing online services: recommender systems and Quality of Experience (QoE).

9.1.1 Recommender systems

The case study of the PersonalTV video delivery service (Chapter 2) showed that the *content retrieval* method (most viewed, top rated, search, or algorithm suggestions) has no significant effect on the *consumption percentage*, i.e. the fraction of the video that is actually watched. A similar conclusion holds for the reported satisfaction of the test subjects. Although a number of trends were identified, the four *content retrieval* types do not yield significant differences in terms of *user satisfaction* with the content, the audiovisual quality, and the way of content retrieval.

Also the relationship between the measured consumption percentage (implicit feedback) and the subjective evaluations (explicit feedback) was investigated. The results indicate that there is no significant correlation between the objective measure *consumption percentage* and two of the satisfaction measures (*satisfaction quality* and *satisfaction retrieval*). In contrast, the *consumption percentage* turned

out to be in proportion to the *satisfaction with the content*: (nearly) complete viewing corresponds to significantly higher satisfaction with the content of the video, thus suggesting a convergence of both measures and implying that consumption percentage could be used as an indirect rating mechanism.

A user-centric evaluation of five different recommendation algorithms (Chapter 3) showed that the Hybrid algorithm, which combines the recommendations of the content-based and collaborative filtering algorithm, outperforms other algorithms (User-Based Collaborative Filtering (UBCF), a Content-Based recommender (CB), a recommender based on Singular Value Decomposition (SVD), and the random recommender) except for the diversity aspect. In terms of diversity, the random recommendations turned out best, which are of course a very diverse set of items. The runner-up for best algorithm in terms of qualitative aspects turned out to be the UBCF algorithm, followed by the CB algorithm on the third position.

An analysis of the relationships between the different qualitative aspects of recommender systems indicated that the *accuracy* and *transparency* are influential predictors of the user *satisfaction*. Because of this effect of the *transparency* of the recommendations on the user *satisfaction*, the low subjective evaluations of the SVD algorithm in the experiment can be explained by the limited transparency of SVD, which has a negative influence on the user satisfaction. An offline evaluation in terms of accuracy confirmed the results of the user evaluation to a large extend, and also indicated that the SVD algorithm is capable of providing accurate recommendations.

To evaluate the effectiveness of group recommendations (Chapter 4), the *aggregating recommendations strategy* (which aggregates the users' individual recommendations into recommendations for the whole group) and the *aggregating preferences strategy* (which aggregates the users' individual preferences into a preference model of the group) were compared. Neither of these aggregation strategies can be designated as the overall winner since the effectiveness of aggregation strategies is influenced by the used recommendation algorithm. Aggregating recommendations is the best strategy in terms of accuracy if individual recommendations are calculated using Item-Based Collaborative Filtering (IBCF) or an algorithm based on SVD, whereas aggregating preferences provides the most accurate recommendations for UBCF or a CB algorithm. Combining the individual aggregation strategies into a new aggregation strategy can significantly improve the accuracy of the group recommendations but increases the calculation requirements.

Furthermore, the influence of the group size and group composition on the effectiveness of the group recommendations was investigated. For randomly-composed groups, the accuracy of the group recommendations decreases as the group size increases. More users in a group means more (potentially conflicting) preferences to take into account for the group recommendations. For groups that are com-

posed of highly similar group members, the results show a higher accuracy of the recommendations. The more similar the group members, the better they can complement each other, resulting in more accurate recommendations. High similarities between group members can even lead to group recommendations that are more accurate than the recommendations for individual users. Besides the accuracy of the group recommendations, also the diversity, coverage, and serendipity were evaluated. Group recommendations based on Collaborative Filtering (CF) have the highest diversity, coverage, and serendipity; the CB algorithm obtains the worst results for these metrics.

9.1.2 QoE analysis

In view of analysing the QoE during mobile video watching, a *controlled environment* (Chapter 6) allows to manipulate the (technical) parameters for the experiment, such as the bandwidth of the data connection used to transfer the videos to the mobile device. These objective technical parameters showed to be highly correlated with the subjective quality assessments obtained via a questionnaire.

A detailed analysis of the influence of these objective technical parameters on the subjective quality assessments resulted in a model for quantifying the acceptability of *video interruptions*. Although video interruptions due to rebufferings are experienced as disturbing, users accept a (limited) number of these rebufferings in a mobile context. Mobile video sessions with less than 10 short rebufferings are in more than 80% of the cases evaluated as 'acceptable'. Furthermore, the subjective assessments of the video quality indicated that the test subjects of our experiment preferred a fluent playback of the video above a higher resolution, frame rate, and bit rate. In comparison with the fluidity of the playback, the test subjects considered the loading time of the video as less critical for having a good experience.

Analysing the QoE during mobile video watching in a real-life, so called *living lab context* (Chapter 7) approximates the real-world that is being examined, thereby allowing to investigate the influence of physical as well as social contextual factors. The results of such a living lab experiment illustrated that most videos are watched at home and in the afternoon and evening. In most video sessions of the experiment, no other people were around during video watching. The presence of other people did not have a significant influence on the overall experience rating and was in the majority of the cases not perceived as a disturbing factor.

Modelling the subjective assessments of the *loading speed* showed that when the loading time increases from 10 to 30 seconds, the subjective assessments of the loading speed gradually evolve from mainly positive to mainly negative. For video sessions using RTP, the *subjectively-perceived distortion* during mobile video watching was modelled as a function of the video packet-loss rate. The probability of receiving a positive rating is rapidly decreasing if more packet loss occurs during video watching (from a packet-loss rate of around 0.25%) and video

sessions with a packet-loss rate higher than 10% are in general evaluated as 'annoying' or even 'very annoying'. Finally, the QoE during mobile video watching can be quantified by a decision tree based on the objective parameters of the video session such as network type and video quality. These results provide application developers and service providers a tool that clarifies which and how technical parameters influence the QoE and how the parameters have to be adapted to optimize the QoE.

Recommender systems and QoE analysis are not independent research domains, since personalized recommendations can influence the general experience of a user with a service or application and because the *QoE* has an influence on the user's *explicit feedback* that is used for generating recommendations (Chapter 8). A living lab experiment with the mobile application of the PersonalTV video delivery service showed that users tend to give a lower evaluation for the content (explicit feedback) if the quality of the video playback is not optimal.

This means that the user's explicit feedback does not correctly reflect the preferences of the user, thereby possibly effecting the accuracy of recommender systems. Modelling the objective technical parameters and the user's explicit feedback enables to correct the user's explicit feedback for video content by considering the influence of a varying QoE. The resulting model can be used by video recommender systems to improve the accuracy of user feedback, and as a result to improve the accuracy of the recommendations.

9.2 Future research

9.2.1 New challenges for recommender systems

Over the last decade, recommender systems rapidly emerged into a tool to increase revenues for online services. Amazon is the best known example of the commercialisation of a recommender system. Amazon is the world's largest online retailer, providing users suggestions on each product page as references to related or similar products, thereby making 20 to 30% of its sales from these recommendations [1]. These commercial recommender systems have to work at greater-than-research scales - handling millions of users and items and hundreds or thousands of transactions per second [2]. As a result, these recommender systems have to face real-world difficulties such as a limited *computation time* and *scalability* issues. Algorithms such as item-based collaborative filtering and dimensionality-reduction approaches are developed to handle these problems, but the integration of more complex algorithms in large-scale commercial systems remains challenging.

Many recommender system operate as black boxes, providing no transparency into the working of the recommendation process, nor offering any additional information besides the recommendations themselves [3]. An increased interest in

a user-centric evaluation of recommender systems [4] has emphasized the importance of user perception, which is influenced by perceived qualities, such as *trust* in the system and *transparency* of the recommendations. *Explanations* can provide that transparency and as a result also trust in the recommender system, by exposing the reasoning and data behind a recommendation [5]. Explanations are important for the end-user as well as for the system owner: e.g., in the context of an online shop, the end-user may look for bargains and explanations that justify decisions, whereas the system owner tries to increase profits by providing convincing arguments for buying [2].

Content-based style explanations are typically based on the item's attributes. A movie recommendation for example, can be explained according to what the system infers is the user's favourite actor [6]. Or a more domain independent approach is to explain the recommendations based on specific keywords or tags that have the recommendation and previously consumed items in common [7]. The most well known example of collaborative-based style explanations are the ones used by Amazon: "Customers who bought this item also bought...". An alternative explanation for collaborative filtering consists of indicating how neighbours or similar users rated the recommended item [3]. More challenging and an interesting topic for future research is the explanation of recommendations generated by less intuitive algorithms, such as algorithms based on matrix factorisation.

Most existing research in the domain of recommender systems focuses on suggesting users the most interesting items based on the user's preferences, but without taking into account any additional *contextual information*, such as the user's location, the device, the time of day, the day of the week, the user's mobility, etc. However, the context is an important aspect in the decision process for the user, particularly for mobile applications. Three different algorithmic paradigms exist for incorporating contextual information into the recommendation process: contextual pre-filtering, contextual post-filtering, and modelling [8]. The contextual pre-filtering approach uses contextual information to select the most relevant user-item data (i.e. users, items, and consumptions) for generating the recommendations. E.g., if a user wants to see a movie on Saturday, only Saturday consumption data is used to recommend movies. The contextual post-filtering approach ignores the context in the recommendation phase, and subsequently adjusts the obtained recommendations using contextual information. E.g., if a users wants to see a movie on Saturday, and on Saturday this user only watches comedies, this contextual information can be taken into account by filtering out all non-comedies from the recommended movie list. In the contextual modelling approach, the context is more interwoven into the algorithm. The contextual information is directly used in the recommendation technique to predict the user's preferences. E.g., probabilistic models can incorporate the context, in addition to user, item, and consumption data, to estimate probabilities. As more contextual information becomes available

through sensors of mobile devices, there is still room for improvement in the field of context-aware recommender systems.

9.2.2 New challenges in the domain of QoE

The user's experience with an application or service varies *over time*: at first use, the user has prior expectations; but during the use process, these expectations may change over time. In addition, also the user's perceptions, personal skills, usage pattern, and motivation may change over time [9]. To take into account the influence of these temporal dimensions and effects, a study with a longer time frame (e.g., one to several weeks) could be set up. Such a study can provide insights into the evolution of the user's experience as (s)he becomes more familiar with the application or service. This can be valuable information for service providers e.g., in order to determine the reasons for user dropout.

In future research, it could be further investigated which *additional factors* might affect users' overall experience and their acceptance or refusal of the produced quality as well as how these factors can be taken into account in order to optimize the experience. These factors can be of a technical nature such as the codec and resolution of the video [10], related to the user and his/her context [11], such as personal skills or the usage environment, and related to the service itself [12], such as the price or the content availability. In this respect, it would be very relevant to also look at other types of mobile devices (for instance smartphones vs. tablets) to see if users adjust their expectations and acceptability thresholds depending on the device characteristics (e.g., screen size).

Finally, the obtained results regarding the quantification of the QoE and the user's acceptability thresholds can be used as input for the development of a *"QoE agent"*. Such a QoE agent can make an intelligent decision regarding the selection of network or content format. For a voice call for example, if a data network with sufficient throughput is available, the call can be made using voice-over-ip. When the data connection is not available or deteriorates during the call, the QoE agent will decide to route the call over the (sometimes more expensive,) GSM network (Global System for Mobile Communications). Another example is the selection of the optimal video format. When the technical conditions of the network and device are optimal, the QoE agent will select the highest-quality version of a video for transmission over the network. Switching to a lower-quality version (lower bit rate, lower resolution, lower frame rate) can be considered, when the technical conditions deteriorate. Ideally, the QoE agent handles this change and selects for each situation the version that will yield the highest experience for the user. In case that the available network provides a very limited throughput, the QoE agent can even decide to change the media type. For live sports commentary for example, the QoE agent can decide to switch to audio-only if (fluent) video transmission over the network is not possible. In case of a network with extreme low

throughput, the audio track can even be replaced by textual reporting. Audio-only or textual reporting as an alternative for video can yield a better experience than no information at all in case of bad technical conditions.

References

[1] B. Yan and G. Chen. *AppJoy: personalized mobile application discovery*. In Proceedings of the 9th international conference on Mobile systems, applications, and services, MobiSys '11, pages 113–126, New York, NY, USA, 2011. ACM.

[2] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.

[3] J. L. Herlocker, J. A. Konstan, and J. Riedl. *Explaining collaborative filtering recommendations*. In Proceedings of the 2000 ACM conference on Computer Supported Cooperative Work, CSCW '00, pages 241–250, New York, NY, USA, 2000. ACM.

[4] P. Pu, L. Chen, and R. Hu. *A user-centric evaluation framework for recommender systems*. In Proceedings of the fifth ACM conference on Recommender systems, RecSys '11, pages 157–164, New York, NY, USA, 2011. ACM.

[5] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor. *Recommender Systems Handbook*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

[6] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. *Justified recommendations based on content and rating data*. In WebKDD Workshop on Web Mining and Web Usage Analysis, 2008.

[7] J. Vig, S. Sen, and J. Riedl. *Tagsplanations: explaining recommendations using tags*. In Proceedings of the 14th international conference on Intelligent user interfaces, IUI '09, pages 47–56, New York, NY, USA, 2009. ACM.

[8] G. Adomavicius and A. Tuzhilin. *Context-aware recommender systems*. In Proceedings of the 2008 ACM conference on Recommender systems, RecSys '08, pages 335–336, New York, NY, USA, 2008. ACM.

[9] D. Geerts, K. De Moor, I. Ketykó, A. Jacobs, J. Van den Bergh, W. Joseph, L. Martens, and L. De Marez. *Linking an integrated framework with appropriate methods for measuring QoE*. In Second International Workshop

on Quality of Multimedia Experience, QoMEX 2010, pages 158 –163, June 2010.

[10] K. Piamrat, C. Viho, J.-M. Bonnin, and A. Ksentini. *Quality of Experience Measurements for Video Streaming over Wireless Networks*. In Sixth International Conference on Information Technology: New Generations, 2009. ITNG '09, pages 1184 –1189, April 2009.

[11] G. Mantovani. *Social Context in HCI: A New Framework for Mental Models, Cooperation, and Communication*. Cognitive Science, 20(2):237–269, 1996.

[12] P. Kortum and M. Sullivan. *The Effect of Content Desirability on Subjective Video Quality Ratings*. Human Factors: The Journal of the Human Factors and Ergonomics Society, 52(1):105–118, 2010.